Spring 1991

Large-scale portfolio evaluation of writing

Jay Simmons
University of New Hampshire, Durham

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

U·M·I

University Microfilms International
A Bell & Howell Information Company
300 North Zeeb Road Ann Arbor MI 48106-1346 USA
313 761-4700 800 521-0600

Order Number 9131297

Large-scale portfolio evaluation of writing

Simmons, Jay, Ph.D.

University of New Hampshire, 1991

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

LARGE-SCALE PORTFOLIO EVALUATION OF WRITING

BY

JAY SIMMONS
AB, Bowdoin College, 1969
MST, University of New Hampshire, 1982

DISSERTATION

Submitted to the University of New Hampshire
in Partial Fulfillment of
the Requirements for the Degree of

Doctor of Philosophy
in
Reading and Writing Instruction

May, 1991

Adviser: Dr. Grant L. Cioffi, associate professor of education

This dissertation has been examined and approved.

Dissertation director, Dr. Grant L. Cioffi, associate professor of education

Dr. Rebecca M. Warner, associate professor of psychology

Dr. Donald H. Graves, professor of education

Dr. Jane Hansen, associate professor of education

Dr. W. Dwight Webb, associate professor of education

Date

DEDICATION

For Cindy, with whom all my dreams are possible

Acknowledgements

Many people have given me the training and assistance necessary to complete this study. Bill Stine dusted off my twenty year-old calculus and taught me statistics, giving unselfishly of his time for a year-and-a-half. Becky Warner kindly agreed to step in when Bill left for Kansas. I am grateful to them both.

I thank Nancy Roberge of Seacoast Educational Services who unravelled scores of administrative snarls with dispatch and good humor.

Jane Hansen has gently forced me to be less rigid about many issues, portfolios included. Dwight Webb I thank for allowing me to learn the lessons counselors can offer teachers.

Without Don Graves this study would not exist. He saw the promise in my original questions about testing practices and has believed in me when I could not believe in myself.

Finally, my fondest thanks go to my advisor, Grant Cioffi. For six years he has encouraged me, taught me, championed my cause, and provided me an example of a caring practitioner who conducts quantitative research. I aspire to be like him.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

LARGE-SCALE PORTFOLIO EVALUATION OF WRITING

by

Jay Simmons

University of New Hampshire, May, 1991

Schools, districts, states, and testing organizations routinely assess writing ability with timed, prompted writing samples. Based on these results, those in control swing the gate open or slam it shut for thousands of students, teachers, administrators and schools, in decisions about promotion, admission, retention or funding.

This study demonstrates that timed writing samples poorly predict actual classroom writing performance, underestimating the weakest and poorest while overrating the strongest. Self-selected portfolios from 263 randomly selected students in grades five, eight and, eleven across eleven school unions also provide a clearer picture of the development of writing abilities from elementary through high school than prompted writing samples.

High school students in this study are less able than their

middle and elementary school peers to predict adult ratings of their work, and less frequently agree with adults about what makes their writing strong. High schools also provide students with the shortest times to develop and revise their writing. Conversely, students seem to progress from grades five to eight in their ability to keep their work open to revision longer and to maintain a sense of adult standards. Finally, students who develop adult-like judgment on the emotional rather than the conceptual or language qualities of papers tend to produce more highly rated papers.

Chapter 1
Changes Needed in Assessment

Shift to Performance Testing No Panacea

Educational Testing Service President Gregory R. Anrig has predicted (Cohen, 1990, p. A33), "There will be more changes in testing in the next 10 years than in the last 50." Many educators and business leaders have seen the limitations of standardized tests focused on a narrow set of skills (Marzano & Costa, 1988; Business Council for Effective Literacy, 1990), preferring instead assessments based on actual student performances (New Hampshire Association of Teachers of English, 1990). "Some 28 states already are administering performance assessments in at least one discipline, most frequently writing" (Cohen, 1990, p. A42).

Performance assessments in writing, generally prompted-essay writing samples, have achieved greater content validity by asking students actually to write. But the focus of the assessment, the written product, may not be the primary focus of instruction in the classroom. Teachers of reading and writing have organized classrooms to empower students to find their own voices, to take charge of territories that provide them with information and enable them to grow in their conscious control of language (Graves, 1983; Hansen, 1987). Evaluations, however, still intend to assess the learning of skills.

Perhaps it is exactly this skills orientation to the evaluation of

writing ability that drives us, so often, to create testing situations that can be broken down into component parts, and to construct evaluative scales that enumerate the contributing factors to writing success (Searle & Stevenson, 1987). This orientation to learning hearkens back to stimulus-response theories of human psychology, stage theories of human development, and production-line models of educational management (Vygotsky, 1978; Apple, 1986; Shannon, 1989). All these frameworks for human activity drive us to superficial measures of normally-distributed human qualities that appear in a predictable sequence. Against this normal curve our society believes we can rate our outlay of capital, both financial and personal.

Variation, however, not congruence, may be the hallmark of writing abilities (Graves, 1983). As a society we may indeed know a good business letter when we read one, but the individual patterns of behavior that create such letters often astound us. Moreover, the value of a business letter, or any good writing, rests more in its effect than its form, and effects are harder to measure and prescribe than forms. Therefore, we have settled too often in the past for measures of written products and called these measures of writing ability. Perhaps it is time to find ways to describe what writers do as they create papers which experienced readers find to be effective.

Assessment as Measurement

Table 1 (p.4) lists concerns of teachers and administrators attending the Meade Conference on the Evaluation of Writing at the

University of Virginia in August 1990. Only #7 fails to mention grading, scoring, counting or rating as the core of "evaluating writing." The first and second refer to scoring criteria for specific genres, implying that some external standard exists or can be validly created against which we can measure the quality of any literary work or art or expository essay. I am reminded of John Keating in Dead Poets Society telling his students to rip the J. Evans Pritchard essay out of their poetry textbooks. Excellence, Keating says, cannot be graphed on the Importance and Perfection axes. "Excrement," is Keating's word for scales to evaluate artistic ability.

Eisner (1990, November) has said measurement "has no inherent connection to evaluation." In fact, measurement is neither a necessary nor sufficient factor in evaluation. Testing, he says, which generally goes hand in hand with evaluation, is merely a shorthand to evoke a response for grading. And grading actually reduces the spectrum of information into a signifier to be matched with a standard.

Genre "standards" are usually set in specific contexts. Editors evaluate work submitted for publication based on "our needs at that time," as many a rejected author can testify. Each publisher sets criteria in the form of an editorial policy. Only the general outlines are consistent.

Consider the scales for evaluating writing in Tables 2, 3, 4, and 5 (pp. 5-8). All four of these mention (in some form): voice, appropriate style, word choice, details, support/development, character of writer, emotion, organization, variety. The four guides

Table 1 Concerns of teachers and administrators attending the Meade Conference on the Evaluation of Writing, University of Virginia, Charlottesville, VA, August 1990.

Conference Topics
1. Setting scoring criteria for different genres, grade ability levels.
2. Grading writing about literature and other subject matter.
3. Portfolio grading and student self-evaluation.
4. Effect of grading strategies on student attitude/performance.
5. How grading strategies influence instructional behavior.
6. How much writing should count in a student's grade.
7. Responding as a dimension of evaluation.
8. Mastery learning and contract grading systems.
9. Departmental reading of students' writing to increase rater reliability.

Table 2 Scoring guide based on Aristotle's Rhetoric, Book III

Scoring Guide

Voice

Natural, clear, appropriate style

strong nouns and verbs

few strange, invented terms

fitting metaphors

easy to read

details

describe instead of naming

express emotion and character

criticism of self

rhythmical to an extent

free-running style

lively and taking sayings

graphic

surprise the reader

connections

organized

focused

supported

Table 3 Scoring guide created April 11, 1990, to rate grade 5 test papers and portfolio selections

SCORING GUIDE

Imagination/Creativity

Move my feelings -- Interesting to me -- ideas, language, viewpoint

Interested writer's voice

Focus -- stick to what talking about

Supporting reasons -- tell me why

Show me not tell me -- Descriptive language/Vocabulary

Mechanics

Logical order -- Beginning, Middle, End

Does form fit what the writing is for?

Appropriateness of dialogue -- amount, place, balance with narrative

Table 4 Scoring guide created April 12, 1990, to score grade 8 test papers and portfolio selections

---

Scoring Guide

Interesting to me

Interesting to the writer

Original/Approach from different pt. of view

Voice

Style

Focus/Zero in on a particular subject, topic, or experience; not bed-to-bed

Varied sentence structure

Use of detail examples

Word choice, varied vocabulary

Dialogue used appropriately

Fluency, coherence

Paragraphing

Organization/Beginning, middle, end

Clarity / lack of wordiness

Grammatical structure / mechanics

Legibility

Table 5 Scoring guide created April 13, 1990, to score grade 11 test papers and portfolio selections

Scoring Guide
Having something to say
Cares, voice of author, honest tone
Style, flavor
Aware of audience
Vibrant
Organization
Unity, consistency
Underlying logic
Grammar, mechanics, usage
Effective use of detail
Development of ideas
Conciseness
Varied, appropriate vocabulary
Use of varied language

were created by (Table 2) Aristotle over 2200 years ago and (Tables 3,4 and 5) by fifth, eighth and eleventh grade teachers in the spring of 1990. The analytic scoring categories of Spandel and Stiggins(1990) seem much the same.

These two authors, and testing programs around the country, would have us believe that analytical scoring of writing tests is the only way to ensure valid, reliable, detailed evaluations of writing ability. Holistic scoring, they say, can only screen or rank order papers (and, therefore, students) and cannot provide the sorts of feedback students and districts need about their writing ability. But we know that the publishing industry follows no set scale and Moore (1985) has found that nearly 90% of all college writing evaluators use no scale at all. Comparison of Aristotle's criteria with those of current teachers should tell us that experienced teachers of writing know what good writing is. That's why holistic scoring works.

Still, Spandel and Stiggins have a point: holistic scoring, reliable and valid as it is, does not provide specific feedback, but Edward White (1985) calls analytic scales "morally bankrupt." White means that analytic scorers either vary so much as to be untrustworthy, as Diederich (1974) found when he devised holistic scoring to begin with, or they form a general impression and adjust the individual category scores to come up with an overall score. Apparently, the need to produce analytic scores in many cases outweighs the ability to do so in a valid manner.

Call for National Standards

The new call for performance-based assessments does not necessarily promise a more valid paradigm. O'Neil (1991) reports, "Support is building for a system whereby national achievement standards would be developed and performances on different tests would be 'calibrated' to those standards" (p.1). O'Neil says the University of Pittsburgh Learning Research and Development Center, directed by Lauren Resnick, and the National Center on Education and the Economy plan "to create national standards to which local, state, or other assessments could be calibrated...(using performance exams, portfolios and projects)" (p.6).

But Monty Neill, associate director of the National Center for Fair and Open Testing (FairTest), in the same issue of ASCD Update, warns, "Even performance-based tests, if misused, poorly constructed, hastily implemented, and made too important, can narrow curriculum and instruction, induce tracking and retention, and penalize low-income and minority students" (p.7).

Broadfoot (1988) characterizes assessment as "descriptive" or "competitive." "Where 'competitive' assessment predominates, the emphasis is likely to be on teaching towards relatively narrow, pre-defined goals," she states. Success in relation to these goals "is reported in terms of grades or a hierarchy of criteria" (p. 296). Such as method clearly echoes the type of "calibration" being pursued by President Bush and this nation's governors. "Descriptive" assessment, on the other hand, Broadfoot says, puts emphasis "on 'testing for teaching' rather than 'teaching for

testing'; where assessment is a component in educational production itself and encourages an educational entrepreneurship among pupils themselves, as much as between schools and teachers" (pp. 296-297).

Opposite Trends Clash in Britain, Too.

Broadfoot reports the British government's move away from externally imposed, standardized testing, toward "records of achievement" in which students are involved in the evaluation process. These records stress "four different aspects of achievement -- knowledge and recall; practical skills; personal and social skills and inner feelings such as motivation, confidence and the constructive acceptance of failure" (p. 292). Broadfoot claims, "To the extent that pupils become involved in the process of target-setting and regular review and are encouraged to reflect on their achievements and development needs, they are being taught the skills of self-appraisal, self-presentation and self-management which are increasingly explicitly being recognised as core skills not only in schools, but in the world of work as well" (p. 292).

Unfortunately, Britain also has more standardized plans, too, "plans which provide for sequenced, staged assessment in most of the main curriculum subjects at ages 7, 11, 14, and 16" (p. 295). She calls these tests "deeply incompatible with the philosophy and purposes of Records of Achievement" because the test results "form the basis of comparison not only of pupils against a given hierarchy of standards but of the standards achieved by individual classes and schools against each other" (p.296).

Societal Bias

For several reasons, educational assessment historically has insisted on measuring these stable, normally-distributed human behaviors (which in writing turn out to be individual papers) on as absolute and differentiated a scale as possible. First, we live in a society that has reified science as knowledge (Apple,1986). Perhaps in reaction to the romantic subjectivism of the nineteenth century, our rationalized habits of social and technological management are based on the assumption that the only trustworthy evidence is experimentally-produced, empirical evidence.

Second, methods available to psychological science, until recently, have only permitted measurement of discrete, external actions. Positivists have measured stimuli and responses in an attempt to combat the philosophical imprecision of mentalists. But as Vygotsky (1978) pointed out nearly 60 years ago, we can no longer allow available methodology to dictate our theoretical questions. Rather, we must formulate theory based on the practices of actual human behavior, in this case the writing habits of writers, and develop the methods to investigate our theories.

Industrial Model

In fact our accepted methods of educational measurement derive from the processing of military recruits in World War I, when thousands of minds and bodies had to be examined and selected, as quickly and cheaply as possible. Emphasis was on fitting a physically and psychologically predictable mold so that uniforms

and training could be duplicated on a grand scale. Newly-developed punch-card technology made it possible and its metaphors still define the data-processing and testing industries today.

At the same time, Apple (1986) and Shannon (1989) point out, an educational management model was being developed based on the production line factory. In mass production the locus of control is external: board of directors to managers, managers to workers. Decisions about the content of actions are made by those furthest removed from the actions. The workers themselves do no managing, only repetitive behavior. Thus, workers can be considered alienated from their work, since they have no control over it. But results, in the form of product or profit can be as predictable as possible to justify the outlay of capital.

These practices translated directly into schools, where boards of education controlled the administration, who controlled the teachers, who controlled the students. Education has been seen as a transmission process, where knowledge, in the form of textbooks and packaged curricula is laid on to passive students. Teaching, in this model, is reified as the application of texts. Writing assessment has become the search for identifiable markers of discrete skills expected to appear in discrete segments of text.

New Demands of Information Economy

Perhaps these assumptions were justified when most school graduates would go to work in production line factories, but current business leaders in Britain (Broadfoot, 1988) and in America (Berger, Dertouzos, Lester, Solow & Thurow, 1989) are calling for actively involved workers who evaluate the nature of

incoming information and their actions on it, and then communicate to others about that information and those actions. In process classrooms, teachers no longer conceive of themselves as applying the knowledge in texts to passive receivers of information, but seek to transfer the locus of control to the writer, and to take control of their own curriculum. The information economy has replaced the production-line factory, and site-based management is replacing the old management model.

Teachers Can Develop Models

Assessment, as far as I can tell, has been the last to change. Now we hear calls for more "naturalistic" methods of evaluation, less standardized, and more in line with the day-to-day activities of teachers at work (Teale, 1988). Teachers, then, need to be given, and to take up, a leading role in the construction of these new models (Glickman, 1990). Glickman argues that the "twin pillars" of the new educational reform movement will be the principle of equal access to education derived in the 1960's and 1970's and that of public accountability of the 1980's. Teachers and schools, he says, will be given the freedom to manage their own affairs as long as they are willing to demonstrate their results.

In Vermont, where they are developing a state-wide system of portfolio assessment of writing, teachers dominate the steering committee. Exemplar schools will develop the process and in-service training will be provided to assist teachers to use portfolios in their own classrooms (Rothman, 1990).

Products or Ability?

Yet, the nature of the results to be reported (numbers, percentages of acceptable or excellent performances per school) is still being determined by legislators and outside policy makers based on outdated assumptions about the nature of learning, reading and writing. Last year I visited with the Vermont state department of education committee. I listened to them discuss logistical issues for two hours, and then was asked to respond with suggestions. First, I needed to ask them a question: "Do you intend to evaluate written products or writing ability?" They looked at me, bemused. The question had not arisen in their year of discussions.

Clearly they wish to evaluate programs that purport to develop writing ability. Teams will evaluate "best pieces" and examples of writing in several genres, as well as read evaluative essays, lists of assignments and logs of conferences. Writing ability, not written products, is the ultimate target of their investigation.

Rosenblatt (1978) says that analysis of a text tells us almost nothing about the habits of the writer who created it. Vygotsky (1978) also points out that product examinations do not illuminate process. Rosenblatt goes further, however, and insists that no "work itself" exists in the text. Works are created by readers as they experience and are guided by the texts. Therefore, each reader is responding to a different "work" in any attempt at "text" analysis.

And what is analytical scoring but text analysis? Are we not attempting to isolate strong and weak qualities of the "text," on

the assumption that certain author skills produced the qualities? Both Rosenblatt and Vygotsky would disabuse us of that notion.

Do we not also assume that if a certain quality (such as correct punctuation) is lacking, then the corresponding skill must also be absent, and, further, that a little instruction in the right place will clear it up? Graves (1983) points out that skills are developed in the pursuit of information, not in reaction to isolated practice dictated by curriculum-based product examinations. Just as product-only tests of writing cannot assess the development of one child's or many children's writing abilities, neither can bigger and better analytical scales isolate the abilities that need work. Writing abilities are neither monolithic nor acretions of lower-order sub-skills represented in text. Rosenblatt (1976) and Gardner (Brandt, 1988) support a search for multiple writing abilities: the habits, preferences and judgments of writers, to be found in the process, not merely the product.

Remediation : the Moral Battle Against Children

Education in America has from the very first been defined in moral terms. New England school masters, in the best Puritan tradition, taught children to read in order that they might read the Bible, thereby fending off the temptations of Satan, the Old Deluder, as Shannon (1989) points out. Left on their own, children would succumb to Satan, not grow to godliness. Instruction was intended to oppose this degenerative trend, was to some degree developed to battle the unchecked development of children.

Even in Piaget's theories of child development (Piaget &

Inhelder, 1969) we can find this vision of battle between instruction and development. Children spontaneously develop concepts, according to Piaget, but ones of such an egocentric nature as to be useless in dealing with the world of adult, non-spontaneous concepts. Before a certain stage of children's maturation, try as we might to influence their thinking, children assimilate our adult concepts into their own egocentric forms. Eventually, we win the battle, however, and in losing, children mature.

Development as Transformation

American education has adopted Piaget's stage theory of development to enhance the predictability of educational planning. Vygotsky, on the other hand, and his work was long unknown to us, demonstrated that growth is marked by periodicity and instability. Graves (1983) has demonstrated that variability, not conformity, is the norm of growing writers. According to Vygotsky's results, higher mental functions are not merely additive piles of lower functions. Rather, as each psychological function is mastered, the whole thinking process qualitatively changes.

Shifting the Locus of Control with Portfolios

External scales and measurement have insisted on a discrete, predictable, additive model of learning in order to control the learning process from the outside. In process-oriented classrooms teachers have put students in charge of their own language. Many current writing teachers encourage students to develop territories

and to become skillful writers and readers as they search for information their readers need. Teachers can also put them in charge of their own evaluation, and portfolios provide the vehicle needed. Portfolios are collections of products, process notes and self-evaluations that enable learners to assess the nature of their own learning, thereby expanding their sense of themselves.

Eisner (1990, November), while down-playing the importance of measurement in evaluation, has called for more perception, selection and reflection in the assessment of complex performances. Gardner and Hatch (1989) have said that portfolios must be examined to assess the development of multiple intelligences. Certainly any evaluation of writing abilities requires the assessment of both complex performance and multiple intelligences. Students who construct portfolios of their best works and the histories of them are making judgments, as well as reflecting on their production and process.

Including Emotions and Attitudes

Sevigny (1981) writes:

> The shortcoming of traditional observation systems is that they quantify through the screen of the observer, and they do not qualify through the screens of the participants. Systematic observers have chosen to ignore the internal states of the participants of the classroom setting. Educational research needs a change in research methodology which would enable classroom investigators to collect subjective data. (p.68)

Sevigny echoes Vygotsky and Rosenblatt's call for literacy research that includes self-reports of the subjects.

But Eisner has also called for the inclusion of assessment items

currently "on no one's list:" critical mindedness, intellectual autonomy, risk-taking, target shifting, participation in a caring community, sensitivity to subtle aspects of life, and a sense of morals and ethics. Here, he clearly emphasizes the role of the learner in a group, not in the isolation of traditional testing, be it standardized multiple choice exams or performance assessment. Corey and Corey (1987) note that statistical measures typically fail to capture the types of learning that result from personal growth groups, and what are writing classrooms described by Graves (1983) and Hansen (1987) if not personal growth groups? According to Corey and Corey group participants who record their experiences and reflect on them demonstrate learning when they refer to goals, feelings of themselves and others, shifts in attitudes towards themselves and the tasks of life, and the need for emotional support of a trusting and cohesive group. John-Steiner (1985), Rose (1989) and Bruffee (1988) have also shown the centrality of group participation in the development of thinking and literacy. Clearly, the time has come for literacy assessment to incorporate the group process rather than assiduously to exclude it.

Product, Process or Program?

Evaluation programs must also decide whether their focus is product, process or program. Often, as we have seen, people evaluate products but intend to draw conclusions about abilities. A portfolio evaluation system developed by Elbow and Belanoff (1986) at Stony Brook formed the basis for pass/re-take decisions in the basic writing course. Students submitted a portfolio of

three pieces written during the course, plus a timed extemporaneous piece. Thus, products were examined, but the decision, presumably, involved whether or not a student had achieved a certain ability level or set of abilities.

Product examinations do not evaluate skill or ability. They simply rate or describe products. If we also ascertain the habits, preferences and judgments of those who produce writing of various levels of quality, I contend, we know more about the skills, abilities, and capacities of performance.

White (1985) says, even if we limit our view to products we should evaluate anything we call writing ability by looking at several pieces or several types produced over a span of time. The Stony Brook system does that.

Gardner (Brandt,1988) says abilities must be measured through examination of production, reflection and judgment. Clearly, we must go beyond the Stony Brook model to determine how our writers look back on their efforts and to find out what bases they have for evaluation.

Product Exams Can Be Improved

Table 6 (p. 22) describes portfolios for three purposes: the assessment of product, process or program. In actual practice, I find these levels to be cumulative. That is, an assessment of ability presumes an evaluation of product, and a rating of program requires an assessment of process and product. Too often, however, evaluations stop at the first stage.

The list of five traits looked for in product examinations is

derived from Diederich (1974) research that led to holistic scoring. A product examination based on portfolios,not timed essays, simply uses these criteria over several pieces produced in class, rather than one written in a test setting. A portfolio evaluation system which rates many pieces of classroom writing merely answers the question, "Which students (or schools, or cities, or states) produce the most highly rated products?"

But portfolio evaluations seem to have two prima facie advantages over test essays. First, they are based on actual classroom writing. While neither product analyses may tell us about writing ability, at least portfolios tell us about the students' actual classroom performance. Simmons (1990) indicates that while test essay scores are correlated to portfolio scores, they only predict 25% of actual classroom performance (R^2 = .25).

Portfolio scores also seem fairer to the least able test takers. My pilot study also indicates that test essays unfairly reduce the scores of only the lowest score group. Thus, those achieving average or better scores on the test do about the same in the portfolio. But those with below average test scores generally achieve average portfolio ratings.

Process or Ability Tests

Seldom, however, do we intend to measure only pieces of writing. Usually we mean to ask, "Has this person developed the ability to leave (or enter) this course (or school)?", essentially the

Table 6 Levels of portfolio information

	Purpose	Traits	Contents
PRODUCT	Assessment of Written Products	Ideas Organization Wording Mechanics Flavor	Best Pieces student choice teacher choice joint selection Required Pieces
PROCESS	Assessment of Writing Abilities	What I do,know,think How I do it, know it... What I feel How I see myself How I approach work How I use/give feedback How I challenge myself	Ordered Pieces student ranking Drafts Notes, journals, logs Labels Conversations Experiences readings teacher notes self and peer evals. conference logs
PROGRAM	Assessment of Development of Writing Abilities	Chances to: +discover topic +use many forms +confer +keep task open +vary length How do students and teachers : +agree on standards +understand the other Change over time Impact on Ability, SES	Assignment lists Conference Logs Interview Data -- +length +duration +range +evaluations +expectations +SES Scores

gate-keeping function of tests. A state proficiency exam in Virginia, the new essay section of the Scholastic Aptitude Test, and the portfolio evaluation in basic writing courses at the University of Louisville (Martin, 1988) all open or shut the gate based on product exams. Clearly, new methods are needed to insure that the abilities they seek to find are actually described. The middle section of Table 6 lists the contents of portfolios intended to measure abilities, complex performances or multiple intelligences. Research must teach us how to use the information we find there.

Program Evaluation: End the Horse Race

As Glickman (1990) has pointed out, schools that wish to retain their freedom and funding in the 90's must demonstrate their results. Therefore, unless authentic programs can be devised, more districts and states will be joining the list of 28 currently using some form of performance assessment, or will fall back on even less valid norm- and criterion-referenced multiple-choice tests. The last section of Table 6 lists the types of descriptive data that Vermont is seeking and that I found to correlate with more highly rated papers in my pilot study.

But the legislative committees and political leaders who commission these assessments and receive the results must be re-educated as well. In the past, they have conceived of program evaluation as a horse race, one in which each group lines up its horses and the ones finishing first win. Last spring when I presented descriptions of the writing habits of high, middle and

low scoring writers in grades 5, 8 and 11 to area administrators, they seemed unable to process the information.

Before them lay results describing how long students worked on pieces in elementary school as opposed to high school, how well different "ability" groups agreed with adult judgments, and what types of writing the most and least able seemed to prefer. The data depicted schools and grade levels where children had fewer chances to use many forms, keep the task open, or develop adult judgment, and these conclusions could easily frame an action plan for the 90's. Yet, before we can form the plan, it seems, we must remove the blinders of the past, a past that has defined education and assessment in industrial, adversarial and moral terms, terms that assume that learners must be controlled from without, rather than invited into the process.

Chapter 2

Review of Writing Assessment and Portfolio Literature

Limitations of single-task models

States and local school districts widely use holistically scored writing samples to measure student writing ability (Cohen, 1990; McCready & Melton, 1981). Some of these tests allow more time or multiple sessions for revision; some do not. Level of difficulty is usually regulated through topic selection. A number of studies reviewed by Breland, Jones, and the Educational Testing Service [ETS] (1984) tested college writers using argumentative or expository topics. Olson and Swadener (1984) describe a similar test of expository writing for the Colorado public schools. In Canada, the Edmonton Public Schools (Searle & Stevenson, 1987) have created achievement tests including writing for years three, six and nine based on prompted, scripted essays. Ratings derived from dictated level of difficulty tests are reliable, but single-task tests fail to examine a range of writing types and may put the least able writers at more risk of failure.

Measurement of more general ability would seem to require more extensive testing. White (1985) prescribes, "We should attempt to measure anything we call 'writing ability' by more than one writing sample and in more than one writing mode" (p. 118). If financial constraints prevent lengthy or repeated testing, he says, we should be particularly alert to problems with validity.

According to Moffett (1981), "The goal of writing through ... a spectrum is not to 'come out on top' but to be able to play the whole range" (p.12). He also stipulates, "Growth means to be able to do more things and to do the old things better" (p.10).

Students may be deemed more mature as writers based on their ability to perform more difficult writing tasks requiring more cognitive development, and to produce effective writing across a range of modes. Both rhetorical and cognitive development theory support the classification of modes of discourse by level of difficulty. Rhetorical theorists have generally defined mode of discourse based on the intellectual task being performed and the intended audience (Kinneavy,1971; Moffett,1981; Ekhardt & Stewart, 1981). Composition professionals have used cognitive development theory to describe writing in terms of the level of abstract thinking required and the distance between the writer and the audience (Lunsford,1981; Flower,1981). Research into the relationship between syntactic structures and mode of discourse confirms that writing narrative requires less cognitive development than writing argument (Crowhurst & Piche,1979).

Limitations of test setting

Since cognitive theory clearly implies that single-task tests would dictate lower scores for less mature writers, who might be able to produce good writing in another mode, the assessment context might be re-designed to generate writing across a variety of modes. Godshalk, Swineford and Coffman (1966) tested

students on a variety of writing tasks -- five essays, three slightly longer than paragraphs and two requiring more complex analysis, and concluded that the increases in reliability that "can be achieved by adding topics or readers are dramatically greater than those which can be achieved by lengthening the time per topic" (p.40). But rather than construct an artificial test made up of many modes of discourse, perhaps we should construct a more natural testing context and analyze results for evidence of performance on a range of cognitive tasks.

Shrock and Foshay (1984) examine testing context relative to professional certification. The objective of such assessments is "to certify that candidates can *perform* core competencies" (p.23) (emphasis in the original), not unlike the aim of school assessment seeking to certify what percentage of students, or which individuals, can or cannot perform certain writing tasks competently. Shrock and Foshay rule out both multiple choice and short answer exams, since if the test-taker's life experience is not reflected in the test question, the writer's ability to respond is impaired. "Most 'answers'", they say, "are thoroughly context bound" (p.24).

Longer essays with more general prompts seem to offer the flexibility writers need. However, Elbow and Belanoff (1986a, 1986b) believe the common proficiency exam "undermines good teaching by sending the wrong message about the writing process: that proficient writing means having a serious topic sprung on you (with no chance for reading, reflection, or discussion) and writing

one draft (with no chance for sharing or feedback or revising)" (p.336). They call for "at least two or three samples of her writing -- in two or three genres at two or three sittings" (p.336). At the State University of New York at Stony Brook teachers of Writing 101 form groups to evaluate portfolios consisting of three pieces in different genres written during the course and a fourth piece done in-class "without benefit of feedback" (p.336).

Martin (1988) has extended the Elbow and Belanoff model at the University of Louisville. Essay exams, she says, did not reflect the staff effort to turn "our faces resolutely away from atomistic 'drill for skill' approaches and toward whole discourse, pressing our students both to read and write texts of increasing length and complexity" (p.30). Martin and the Louisville students, as had Elbow and Belanoff, "saw that the exit exam devalued their entire semester's work" (p.31). Instructors also felt they spent too much time devising prompts and scripts for the exams, time that could be more effectively spent on portfolios, "deciding how to evaluate these diverse works...to enhance the sense of collaboration and community among teachers and to extend the developing sense of collaboration among students" (p.32).

Greenberg and Witte (1988) also cite arguments for increased validity in direct writing assessment. Greenberg calls for multiple samples collected by teacher-created instruments. Witte says that since writers engage in different tasks to complete different prompts, "no one prompt will assess ability" (p.13).

Anastasi (1982), on the other hand, cites several studies in support of her claim that "extraneous factors are more likely to

operate with unstructured and ambiguous stimuli, as well as with difficult and novel tasks, than with clearly defined and well-learned functions." Children, she says, are more likely to be affected than adults.

White (1988) makes similar claims for essay exams, calling for reduced sources of variability, constant criteria, pre-tests, control of prompts, control of reading and scoring procedures, and multiple measures. Spandel and Stiggins (1990) claim that analytical, not holistic scores, provide reliable, useful scores.

The need to measure more than products

Brandt (1988) reports that Howard Gardner of Harvard's Project Zero insists that talent must be measured through production, perception and reflection. Portfolios can more effectively address these demands than conventional testing, Gardner says, that is often done in quest of "isolation." Gardner asks what could be more isolated than a portfolio of finished and rough work, diaries, commonplace books, and the evaluations of the student and peers. Hatch and Gardner (1986) suggest measuring, perseverance, flexibility, self-awareness,as well as persistent interests and special abilities in order to understand a young child's giftedness. The methods, however, they feel are also applicable to ordinary students in everyday classrooms, when ability is being measured.

Valencia et al. (1989), discussing reading assessment, expand the scope of traditional measures in that area as well. They argue that real-life contexts in which to apply skills, a positive attitude

toward the self as a literate person, the process of constructing meaning, and attitudes, habits and self-perceptions associated with language use must be included in assessment.

Burnett(1985), discussing the use of portfolios for the purpose of granting college credit for life experience, points out that assigned essays are generally modeled on the professional article, where topic, length and preparation time are limited by outside sources. Personal narratives of learning (the self-reports suggested above) "more closely resemble the experience of learning" (p.45) In these narratives, Burnett finds evidence of learning in students' descriptive statements about products, analytic statements reflecting awareness of process, and interpretive statements reflecting changes in attitude and the setting of goals.

Flood and Lapp (1989) also call for more than a single test score to reflect language ability in reading and writing. In addition to formal measures, they call for informal assessments of writing development; self-reports; change over time; awareness of audience, voice, organizational skills, thinking, language mechanics; the choice and number of books; and the number of minutes read. These last items stress the length and duration of a learner's work.

DURATION

Getzels and Csikszentmihalyi (1976) assert that, "If creativity is to be understood, examination of a finished product is not enough" (p.5). They wish to examine "the formulation of a creative

problem to which the solution is a response" (p.5). When artist-critics, artist-teachers, business executives and mathematicians evaluated works of art, the "problem-finding" referred to above, they found, "related more to originality than to technical skills" (p.121).

In fact, "artists who defined their problems soon after starting work produced drawings that were less original than those who kept the problem open longer...Delay in closure helps to insure that the artist will not settle for a superficial or hackneyed problem" (p.247).

Of course, artists and art schools have long created and evaluated portfolios designed to reflect ability through a range of products. Getzels and Csikszentmihalyi (1976) hold that "in principle the same problem-solving paradigm seems to apply to forms of creativity other than the artistic" (p.248). Therefore, a writing portfolio measures student ability to create original solutions to writing problems over a longer duration than a writing sample permits.

NATURAL SETTINGS

Teale(1988) advocates using "the means of assessment that help teachers in their day-to-day instructional decisions" (p.175). He calls for accounts of student behaviors, checklists and inventories, and "structured performance samples"(p.174), to fit assessment to the learner and the teaching situation. Although Teale is not specific, "structured" seems to imply some range of task in the samples. Developmental information can be charted and summarized, he says.

Faigley, Cherry, Joliffe, and Skinner (1985) recommend writing ability be evaluated by a combination of ethnographic studies, text analysis, and verbal reports from writers. Such a method, while admittedly expensive, would evaluate program as well as ability.

Della-Piana et al. (1988) have attempted to assess growth in writing by externalizing aspects of the reader-writer conference over particular papers. They stress the need to assess process as well as product, in particular the types of writer responses that follow certain reader responses. This work implies that a sense of audience and responsiveness to it in the writer's on-going dialogue with the self should correlate with better written products. Also, students who match teacher perceptions about the qualities of their papers should receive better scores.

EMOTIONS

Here again, the counseling literature can be instructive, insofar as the development of writing abilities is personal growth. Boy (1990) asserts, "Clear thinking emanates from a person whose feelings are known and under control. Faulty thinking emanates from a person who is unsure of how feelings influence our ability to reason well" (p.4). "When examining any behavior and attempting to determine the cause of that behavior, we must avoid examining superficial causes...the deeper cause for the behavior is the person's feelings about these conditions" (pp.5-6).

Boy suggests that counselors must reflect clients' feelings in order to help the clients change. Specifically, reflecting feelings allows "the locus of evaluation to be in the client" (p.14), gives the

client the power of choice, clarifies the client's thinking, and develops a positive attitude about the self. Finally, the ability of counselors to reflect feelings allows them to avoid three common pitfalls: being an analyst, being solution-oriented, and not allowing the time for clients to identify the more serious problems and to solve them themselves. Therefore, to the extent that teachers and students agree about the emotional content of their writing experiences, development of writing abilities should be enhanced.

Components of Portfolios

A VARIETY OF DRAFTS

Portfolios for high school drafting students, art school graduates, second language students, minority high school students, technical writers, and public-schoolers contain many of the same elements. Henderson (1982) has created a drafting curriculum for the Greenville, SC, schools, including portfolios used to assess students moving to post-secondary instruction. Corrected and non-corrected work, drawings suggested by the instructor as articulated in the curriculum, and optional pieces highlighting areas of interest to the student would be included.

Lammon (1985) suggests fine artists seeking jobs use rough drafts, labels, and notes discussing choices during creation to indicate their professional artistic instinct. They also should include areas of special interest to themselves, as well as a range of work to show versatility and to avoid being stereotyped. She emphasizes demonstrating the speed with which work can be

completed.

McLean (1987) advocates measuring second language achievement with cumulative folders of performances in diverse situations over the course of several years. Children would collect the work, both corrected and rough pieces, including efforts from the beginning to the end of any year. Such a method, already required by the Ontario Ministry of Education, has more pedagogical validity than mere sorting and ordering, especially when care is exercised in the design and survey of the item pool, he says.

SELF-EVALUATIONS

Cooper (1990) details 12 components of portfolios for minority students preparing for college in San Diego, CA. These include: a timed writing sample, in-class essays, a best set of notes, a learning log, writing done in another academic class, annotated college research, long and short term goals and plans, a personal evaluation of growth, an example of sudden insight, evidence of community service, a "wild-card" example of accomplishment causing pride, and a synthesis of the year's work.

Bishop (1987) in a technical writing program uses three levels of drafts: rough, professional, and portfolio (for a grade), as well as self- and peer-evaluations in her mid- and end-of-the-year portfolio evaluations. In Bishop (1989) she uses a check-list of class activities, including a literacy autobiography and interviews with professional writers, as well as a rubric for A,B,C,D or F portfolios to see if her assessment matches the student's rating. In 90% of the cases it does, she says.

In the Pittsburgh Arts Propel Project with public schools portfolios consist of biographies of works, range of works and reflections (Wolf, 1989). The portfolios focus on student judgments and choices that cannot be traced in products alone, "not to be found in the text, but in thinking back to earlier times, comparing pieces, and struggling to put your intuitions into words" (p.38).

BRIEFER SAMPLES

Mathews (1990) describes reading/writing portfolios in use in the Orange County (FL) public schools. Four required core elements are : a reading development checklist (completed by the teacher watching the student),writing samples, a list of books read, and a test of comprehension. Teachers or students could choose to include self-evaluations, reading records, anecdotal records, or pages from logs.

Killingsworth and Sanders (1987) suggest professional communications majors construct a job placement portfolio of finished work demonstrating quality, variety, professionalism and maturity. They and Ware and Jewell (1988) also suggest a pared-down, non-returnable sample from the portfolio be mailed to the prospective employer.

Pearson (1988) has devised an album assessment of reading development in which an exercise book is used to collect comments, responses and samples of a student's work over the course of a year. The class teacher, head teacher, other teachers, adult helpers, parents and the child may contribute observations about attitudes, responses and strategies displayed by the child

during reading.

Exit Exam Alternative

Martin (1988) follows Elbow and Belanoff in using portfolios to replace the exit examination from a college basic writing course. Her staff asked for a personal essay, an essay based on outside experience, and an in-class writing. They found they could score diverse pieces of writing reliably on a four-point scale (high pass, pass, barely fail, abjectly fail), that knowledge of the original assignment was unnecessary, as was the in-class piece. Although they processed the longer papers quickly, they experienced fewer discrepant scores than had occurred with prompted essays, had a slightly higher rate of passing, and found more risk-taking and authentic writing in the portfolio pieces.

Program Evaluation

In Vermont schools will be responsible for the collection and maintenance of portfolios for all students in grade 4 and 8. Teams will evaluate "best pieces" and examples of writing in several genres, as well as read evaluative essays, lists of assignments and logs of conferences. Evaluators will attempt to assess how frequently students observe mechanical conventions, use suitable organization, exhibit personal expression, use clarifying detail. Evaluators will also show how frequently programs allow progress from early to late drafts, variety to challenge all learners and provide success, teacher and peer response to drafts, and

opportunity to revise.

Problems in Portfolio Use

STUDENT CHOICE

Student choice provided for by portfolios may create additional benefits. Freedman (1983) notes that while students "see themselves performing well on difficult yet interesting writing tasks" (p.322), they don't often find such tasks. Rather, "those that are difficult are dull, and those that are interesting are easy" (p.322). Forced-choice tasks may produce duller writing.

According to Comstock (1988) fifth-grade students in Stratham, N.H., claim their "own writing" (p.1) will tell how well they write. If the evaluators simply looked in the writing folders, or, better yet, had the writers select their own best work, the evaluators would learn more, the children say. Assigned topics, moreover, can be "too personal" (p.1) or prevent you from showing "how good you are, if you don't like the topic" (p.2). Their comments confirm concerns of Gardner and others: time limits force them to drop good ideas unsuitable for short pieces, give them no time to think, and allow them little chance "to check skills, proofread" (p.2).

In my pilot study (Simmons, 1990) range of modes of discourse positively correlated with portfolio and test score. That is, students who chose a wider variety of modes for the portfolio scored better, and, therefore, the ability to value and perform a variety of types of writing seems connected to the production of more highly rated products. Curtailing student choice fails to

provide this measure of judgment.

According to Hyde and Linn (1988) there are no gender differences in verbal ability at any age level. They base their conclusion on a meta-analysis of numerous studies, finding a weighted mean effect size so small that they consider gender differences in verbal ability to have disappeared. Simmons (1990), however, reports girls significantly outscored boys in student-selected portfolios and prompted essays. The differences may be due to the small sample size, but standard for choice may play a role. Graves (1975) notes "unassigned writing may give a more valid developmental profile" (p.9), since boys and girls differ in their use of territory and first person. Gilligan (1982) posits that males and females conceive of story differently. Therefore, the match between the gender of the rater and the gender of the writer may also play a role in determining paper score.

Overall, we might expect student criteria to vary widely from those of adults. Newkirk (1984) found, writing for peers (which a student-selected portfolio arguably is) may fail basic academic requirements, since students seem to use criteria widely different from those of teachers. Results of Simmons (1990) also showed that while students most often list flavor of the piece or experience of the writer as the strength of the writing, teachers tend to choose ideas or organization. Yet, the students who most closely matched teacher patterns of judgment (omitting flavor and experience in favor of ideas and organization) achieved the highest scores.

Shrock and Foshay (1984) reject portfolios for professional

certification assessment because of problems related to choice. First, certain organizations might not permit inclusion of relevant pieces, or the work situations of some might not be conducive to producing certain types of valuable portfolio pieces. Second, authenticity might be hard to assure, and the contributions of others in the workplace might be hard to separate.

For the purpose of program evaluation, the first concern of Shrock and Foshay actually argues for the use of portfolios. Programs by district or grade level that do or do not encourage the production or inclusion of particular forms of discourse would be identified by such a process. As for the second concern, Simmons (1990) found that rankings based on isolated prompted essays were significantly correlated with rankings generated from ratings of self-selected writing produced in the classroom over a period of months. Therefore, the concern with isolation and authenticity would seem unnecessary.

COST OF ADMINISTRATION

Mary Fowles of the Educational Testing Service reports portfolio assessment of third-graders' writing in rde Island has attempted to measure "a myriad of skills that come into play as writers approach different tasks..." (p.13), evaluative judgment being one (Benderson, 1989). Students wrote papers in a variety of modes and "letters to the project leaders evaluating the assignments" (p.13). However, as Ines Bosworth says of the rde Island project,"Big bucks are involved in portfolios, but they are more valid than a single-shot assessment" (p.14). Since Simmons

(1990) and Martin (1988) report scoring of portfolio pieces can be done quickly and reliably, portfolio assessment of random subsets of the population may prove as cost-effective as global essay testing.

PROBLEMS USING PORTFOLIOS FOR PROGRAM EVALUATION

Kemp, Cooper and Davis (1990) warn that commercial ventures in portfolio assessment threaten to offer busy administrators slick packages, but that early portfolio failures seem due to teacher resistance. They suggest involving teachers early and often in the planning, construction and implementation of portfolio projects.

Discussing effective program assessment for adult literacy, the Business Council for Effective Literacy (1990) prescribes using multiple instruments intended to address the goals of the learner, focusing on the experience and strengths of the learner, not being imposed from without, nor being separated from the regular course of learning, but being a collaborative effort among "the teacher, the learner and the text" (p.7). They suggest interviews, interactive readings, portfolios of writing, observation, simulations, and demonstrations. They note that "a major task confronting the field is to systematize alternative assessment approaches into strategies that can be used in a wide range of contexts" (p.8). They also quote Susan Lytle of the University of Pennsylvania Adult Literacy Evaluation Project, who holds that the workplace may lead the way in alternative assessment development since there assessment is tied most closely to the

"meaningful use of literacy in a context" (p.8).

Wolfe (1989) sounds a more pessimistic note. "'Alternative assessment' has been defined as essentially program-based and learner-centered and as involving a range of procedures which together provide a rich portrait of individuals' learning over time. With this definition it is not possible, therefore, to imagine a city-wide alternative assessment" (p.3). She also calls for efforts "to determine which features of the writing in portfolios count as evidence of change and growth" (p.4), but she goes on to call for defining the choices students should have in the assessment process, determining which records and interview data are useful and how to present the results to outside reviewers and funders.

Buddmeier and Raivetz (1990) report efforts to flesh out the results of required, prompted essay exams with portfolios in New Jersey, a state that has mandated a "high-risk, statewide graduation test" (p.1). They have collected prompted essays from the same students in the third and sixth grades, and a random selection of them in grades four and five. Scoring them on a holistic scale with the same training packet year-to-year, they have found the scores moving up.

Teachers have been sorting student papers to "reconstruct portfolios" (p.8) made up of the large-scale prompted pieces and one teacher or self-selected piece from another year, plus data sheets of interview and teacher input and reviewer responses. The prompted papers have shown consistent growth across years, but the folders have been inconsistent.

In particular, the project has found prompts affect the quality of the essays, the reviewers have difficulty using a common terminology to discuss change, and enormous time is required to review the folders. Reviewers in the New Jersey project, as well as evaluators in Vermont, examine individual student folders to determine histories of change and environments that affected the attitudes, practices and purposes of the student writers.

Portfolios Demonstrate Learning from Life Experience

While few projects exist to use portfolios systematically to evaluate writing ability across large populations, efforts abound in the post-secondary education and professional certification fields. Numerous institutions report using portfolios to assess learning in settings less structured than the traditional classroom. Typically, these projects ask students applying for course credit for "life experience" to submit a portfolio and explain how they have met specific course requirements.

Burnett(1985) identifies the types of statements in student narratives that demonstrate learning: descriptions of products, interpretation of knowledge, analysis of changes in attitude and insight. Dagavarian (1989) says 90% of applications were successful at Thomas A. Edison State College. Students followed course descriptions, discussed relevant theory, mentioned their training and work experience, and included evidence in support of their requests.

Barba et al. (1985) describes a similar system for the placement of candidates into nursing programs. Portfolios as

"alternative transcripts" (p. 121) require applicants to connect evidence of learning to their goals, required course competencies, theory and practice in the field, and current course content.

Clearly, portfolios require credit applicants to use higher order thinking skills, or to think like professionals. Marsh and Lasky (1984) mention three key cautions with portfolios seeking such a commitment of energy from students. First, the narrative demonstrations of learning may be difficult for students unfamiliar with professional and academic environments. Second, with considerable emotional investment in the product...failure to receive credit may become personalized and may be more devastating for some than failure of an examination" (p.267). Finally, the authors echo Wolfe and Lytle, asking how reliability, validity, and resource needs will be met. Jesser (1984) describes a career portfolio process for high-school students in Colorado.

Portfolios for Teacher Certification

Despite the preference of Shrock and Foshay (1984) for simulations over portfolios for professional certification, at least four studies report such efforts. Geiger and Shugarman (1988) report use of portfolios maintained over the undergraduate years as evidence of professional decision-making. Williamson and Abel (1989) suggest tailoring the portfolio to the job sought. Terry and Eade (1983) describe a plan in which the prospective Florida teacher collaborates with a support team to reflect on the individual's strengths and needs, collect data and commit to goals.

McLarty et al. (1985) describe a multiple-measure system for identifying excellent teachers deserving of career rewards. Portfolios were used in conjunction with observations, interviews, a test, evaluation, and questionnaires. Each part was felt to measure important but separate areas of competence, since the measures correlated poorly with each other. The overall system was rated effective but cumbersome.

Purpose of Current Research

In this study I asked whether timed tests produce different estimates of student writing ability than scores based on self-selected portfolios of students' best work. If the scores are correlated, which are higher, and which students fare better in which context?

I also asked whether portfolio assessment of a random subset of students yields more information about writing abilities across a number of school unions than data produced by holistically-scored extemporaneous tests administered globally. Do certain modes of discourse lower scores of fifth-, eighth-, or eleventh-grade writers? Do the duration of work, length, and time of year of self-assigned writing characterize writing by fifth-, eighth- or eleventh-graders of varying abilities? Do students who submit more varied types of writing receive higher grades? Do girls write differently from boys? Do male and female raters react differently to writing produced by girls as opposed to boys? Do populations from different socio-economic levels produce

different writing profiles?

Finally, I attempted to measure the degree to which student judgments about portfolio writing match those of their teachers, and to decide what these patterns of association tell us about writing ability.

The school consortium for which I worked collected and scored portfolios of a random subset of the population in about the same time required for global test-sample assessment. If portfolios generate richer, unbiased information about a school's writing ability, this study provides a powerful model for large-scale writing assessment.

Chapter 3
Method

Planning

Planning for the 1989-90 Writing Sample began in the spring of 1989 when teachers returned questionnaires approving a March/April sampling date and a general prompt. In the fall representatives of participating school administrative units (SAUs) met with me to be briefed on the portfolio collection process and to offer suggestions to aid collection of a prompted test piece and three student-selected "best-pieces." Teachers asked to be notified early in the year which students would be selected so that collection could begin early. The superintendents later decided against early notification to prevent skewing the sample or putting too much pressure on selected students.

Subjects

Participating SAU provided class lists to the research office, where each of the 4016 students was given an identifying number. I constructed a stratified random sample such that each SAU contributed to the sample proportionately to its representation in the overall population. I used a random number chart to select numbers of students who were asked to participate. At least seven students were selected from each school union. Students assented and parents or guardians consented to each student's participation.

Collection

Schools were notified through the superintendents and principals that collection would take place between March 15 and April 1 from students to be randomly selected from class lists (see Appendix A). Names and complete instructions were sent to teachers with the test booklets and portfolio coversheets (Appendix B) in the first week of March (see Appendix C). Portfolios began arriving in the research office after April 2. The last portfolios, a batch from grade 5, arrived a day after the grade 5 scoring session April 11. High school papers were often graded classroom assignments, whereas grade 5 pieces frequently arrived with covers, artwork and text photocopied to preserve the valued originals.

I collected 263 portfolios submitted by students in grades 5, 8 and 11 from 11 SAU in the area: 115 from grade 5 (76% participation), 87 from grade 8 (71% participation), and 61 from grade 11 (46% participation).

Procedures

SAUs were rated by socio-economic status (SES) based on free and reduced lunch counts. The number of free lunches per sampled SAU plus half the number of reduced lunches per SAU was divided by the gross number of students per SAU. The resulting ratio was converted to a whole number, so that the higher the number, the lower the SES.

In order to examine the effect of student choice in portfolio

construction, we had each subject select three best drafts done during the 1989-90 school year that would show how good a writer he or she is. Each writer noted on a coversheet (Appendix B) the month of the school year during which the piece was begun and the month during which the final draft was finished, as a gauge of duration of work.

To measure students' evaluative thinking, I asked for three qualities each felt made individual pieces good enough to be in the portfolio.

As a test of the effect of writing context, subjects wrote a fourth sample during a common one-and-a-half-hour writing period. Dictionaries were available. Teachers devised a general writing prompt intended to allow response in many modes of discourse: "Write about something you know and care about. Make sure your reader knows how much you know and how much you care." Based on Simmons (1990), I predicted raters could evaluate pieces in many modes as reliably as test papers restricted to a single type of writing.

A complete portfolio consisted of:

1) a test piece

2) the mode of discourse of the test, the writer's gender, the genders of the scorers and their ratings

3) three pieces chosen by the student to show how good a writer he or she is

4) a coversheet for each, listing: mode of discourse, three reasons why the paper was chosen, month the work began, duration of work, length of the paper, the student's own rating (2-8) of the piece, the (2-8) rating the student would expect from a teacher.

In the spring in three separate sessions, 16 scorers met at an

area junior high school to score the fifth-grade sample, 15 eighth middle/junior high school teachers scored the eighth-grade papers, and eight high school teachers read the eleventh grade sample. Raters processed papers at about 6 minutes each.

Teams were trained prior to scoring to achieve reliability (Peters, 1981). Although teachers did not score papers of their own students, each paper (with the identity of the writer masked) was otherwise randomly assigned to two scorers, who were blind to the study's hypotheses. Paper score was measured on a 2- to 8-point holistic scale; each rater assigned a score of 1 to 4, and those scores were added. Discrepancies of more than one score level were resolved by a third rater. Tables 3, 4 and 5 represent the scoring guides generated by the three scoring groups. Charts 6, 7 and 8 list the qualities of papers scored 1, 2, 3, or 4, as they were recalled by scorers at each grade level.

This method duplicates writing sample scoring in the region over the past ten years. Interrater reliability was estimated by Cohen's Kappa to be: grade 5 --.99 test, .97 portfolios; grade 8 -- .95 test, .98 portfolios; grade 11 -- .97 on both test and portfolios.

Raters listed three strengths of each paper at the time of scoring, so that I could compare their criteria with those of students. Raters also indicated their gender, so that I might test whether the match of gender of reader with gender of writer has any effect on score. Gilligan (1982) says that males and females conceive of story and character differently. Girls in my pilot study (Simmons,1990) significantly outscored boys on both test and portfolio measures.

After the scoring sessions, seven experienced teachers of

writing were hired to sort the student comments into categories used by the teachers, groups based on Diederich (1974): ideas, organization, wording, mechanics, and flavor. I also included "experience" (referring to the experience of the writer in writing or sharing the piece), in an attempt to capture the student tendency to focus on experiences surrounding the piece of writing, but not included in it, as reported by Newkirk (1984) and Benedict (1989). Interrater reliability for this sorting was .69, as measured by Cohen's Kappa.

I hypothesized that students capable of greater range of writing will achieve higher scores, as they did in my pilot study Simmons (1990). Experienced teachers of writing at each school sorted test and portfolio papers into the following mode of discourse categories: narrative, descriptive, expository, argumentative, poetic. High school students counted the number of different modes in each portfolio, yielding the range score (1-3).

Chapter 4

Results

Effects of Assessment Context

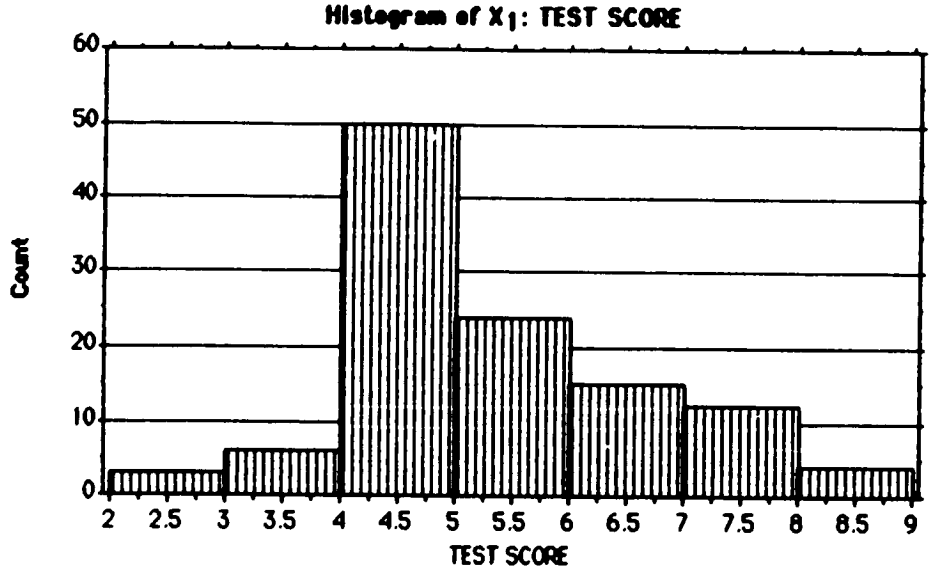
Kolomogorov tests of portfolio and test scores find non-normal distributions of both sets of scores at all grade levels. Histograms (Figures 1, 2 and 3) indicate different shapes for the distribution of test and portfolio scores, as well. In grade 5 (Figure 1, p. 52), although the ranges of the distributions are equal, as is the skewness (test = .558, portfolio = .534), the test curve is slightly flatter (kurtosis: test = -.093, portfolio = -.079), and the test curve has fatter tails.

In Figure 2 (p. 53), the grade 8 test score curve skews more to the right (skewness: test = -.004, portfolio = -.093) and is slightly flatter (kurtosis: test = -.653, portfolio = -.301). Test scores also range from 3 to 8, while portfolios range from 2 to 7.

A reversed picture appears in Figure 3 (p. 54). Here, grade 11 test scores extend further left (skewness: test = .439, portfolio = .934), and the portfolios extend over the shorter range. The test score curve, again, appears to be more platykurtotic (kurtosis: test = -.2, portfolio =.242).

Because test and portfolio scores are distributed non-normally

Figure 1 Histograms of grade 5 test scores (mean = 4.82) and portfolio scores (mean = 4.77)



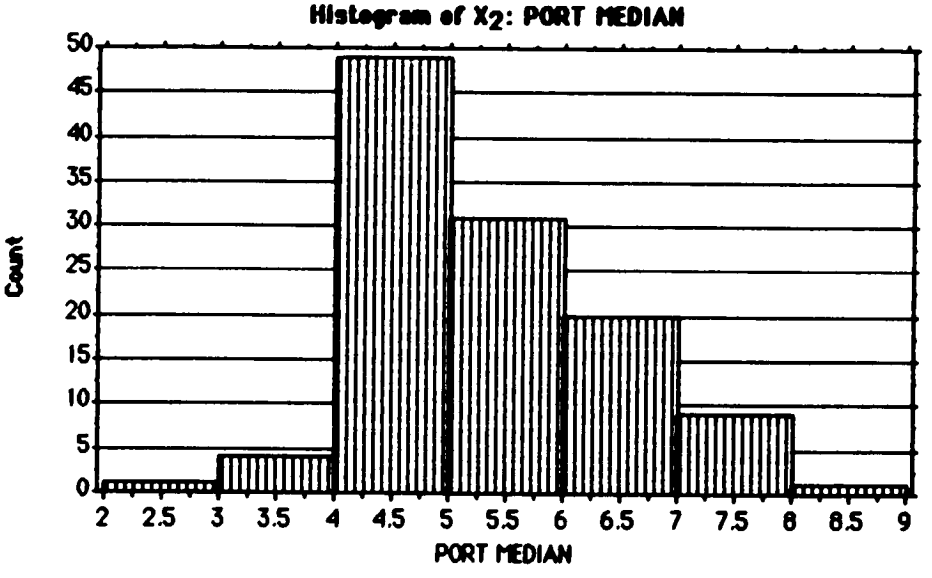
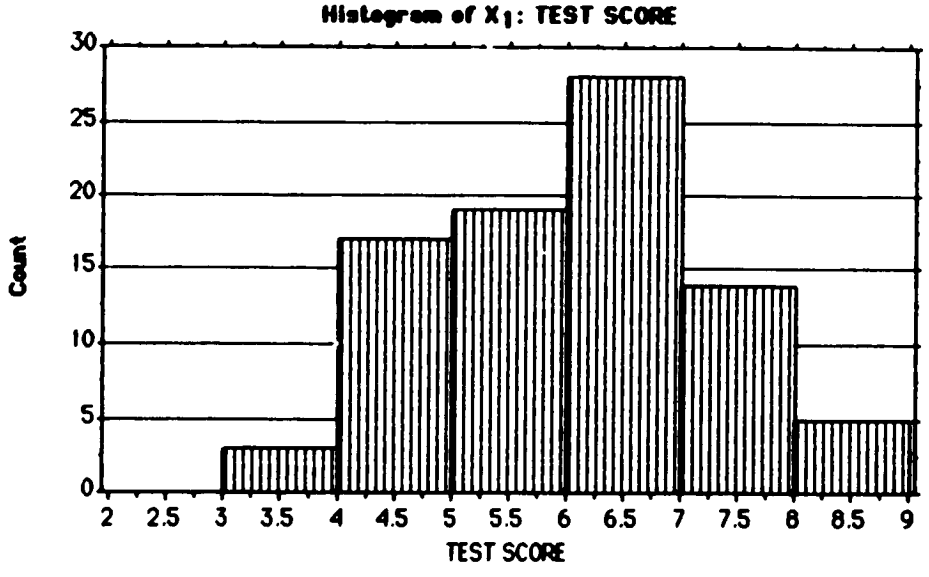


Figure 2 Histograms of grade 8 test scores (mean = 5.55) and portfolio scores (mean = 4.64)



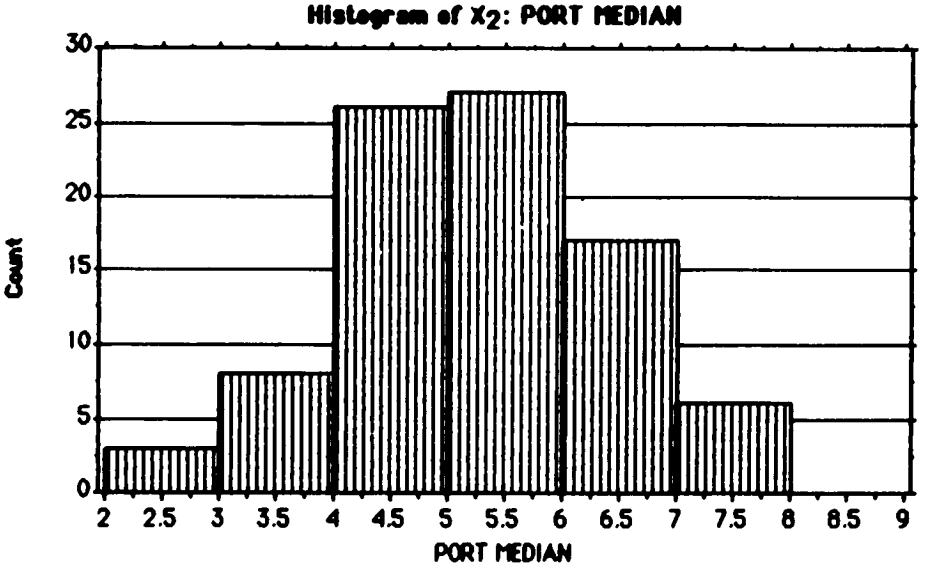
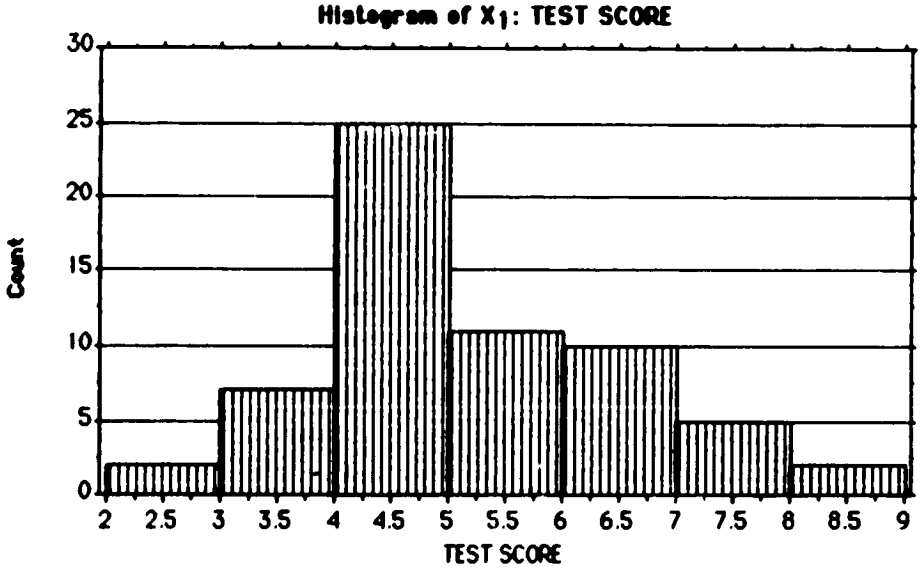
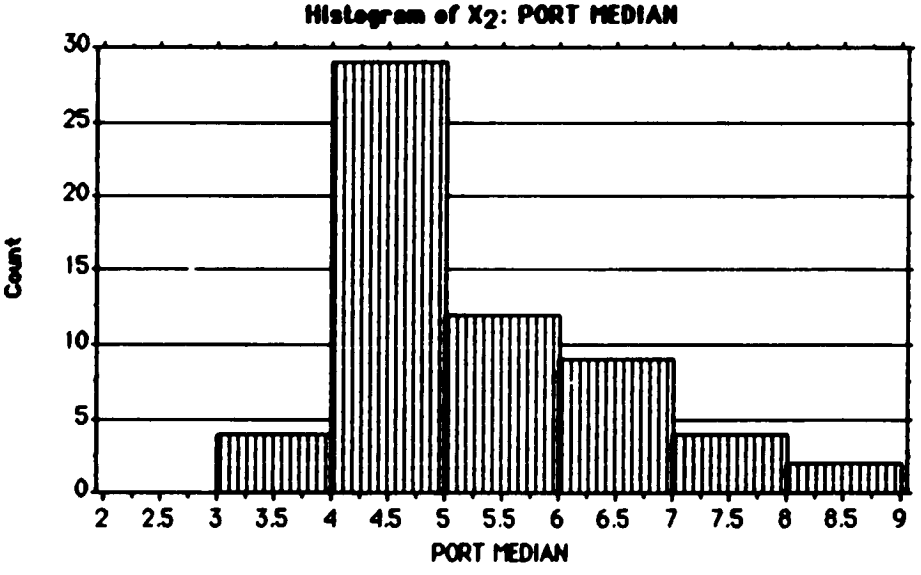


Figure 3 Histograms of grade 11 test scores (mean = 4.72) and portfolio scores (mean = 4.79)





and involve an ordinal scale, non-parametric statistics are appropriate. Since distribution shapes vary, comparison of score groups is undertaken through Chi-Square contingency analysis of test score group membership compared to portfolio score group.

According to Spearman tests of rank correlation, test and portfolio median scores are significantly correlated at all grade levels, as shown in Table 7 (p. 56). Sign tests indicate that test scores significantly surpass median portfolio scores in grade 8 ($z = -4.644$, $p < .00003$) but not in grade 5 ($z = -.187$) nor grade 11 ($z = -.453$).

Chi- Square contingency analyses of the association between test score group (low = 2 or 3, middle = 4 or 4, high= 6, 7 or 8) and portfolio score group (low, middle or high) find significant associations between test and classroom performance at all grade levels and in the overall population, as shown in Table 8 (p. 57). Examination of observed frequency tables (Table 9, p. 58) indicates that 2 0f 9 (22.2%) low test scorers (column 2) in grade 5 also scored low in classroom writing chosen for the portfolio, while 7 (77.8%) scored average or above. In grade 8, 2 0f 3 low test scorers (66.7%) also received low portfolio scores, but 1 received a middle score (column 2). In the high school sample all 8 (100%) of the low test score group placed in the middle (87.5%) or high (12.5%) portfolio groups. In the total sample, only 4 of 20 (20%) received low scores on both test and portfolios, while 16 (80%) fell in the middle or high portfolio groups.

In the low portfolio row, we can see that in grade 5, 3 (column 3) or 60% of those scoring below average in classroom writing had achieved middle test scores. In grade 8, of the 11 lowest rated

Table 7 Spearman rank correlation of test score and portfolio score

	Grade 5	Grade 8	Grade 11
Number	114	86	60
r	.523	.273	.435
z	5.563	2.52	3.341
p	<.0003	<.006	<.0005

Table 8 Chi-Square measures of association of membership in low, middle, or high test score group with membership in low, middle, or high portfolio score group in grades 5, 8, and 11, and in the total sample.

Grade	df	X^2	Cramer's V	p
5	4	29.10	.357	<.0001
8	4	14.37	.289	<.0062
11	4	17.40	.381	<.0016
total	4	40.56	.280	<.0001

Table 9 Observed frequency tables of test score groups (columns) and portfolio score groups (rows) in grades 5, 8, and 11 and across the entire sample

Grade 5

	Low	Middle	High	totals
Low	2	3	0	5
Middle	6	60	13	79
High	1	11	18	30
totals	9	74	31	114

Grade 8

	Low	Middle	High	totals
Low	2	6	3	11
Middle	1	25	27	53
High	0	5	17	22
totals	3	36	47	86

Grade 11

	Low	Middle	High	totals
Low	0	3	1	4
Middle	7	29	5	41
High	1	4	10	15
totals	8	36	16	60

Total Sample

	Low	Middle	High	totals
Low	4	12	4	20
Middle	14	114	45	173
High	2	20	45	67
totals	20	146	94	260

portfolios writers, 6 or 54.6% (column 3) had middle test performance and 3 or 27.3% (column 4) had scored high on the test. In the high school sample, all four low portfolio writers scored either middle (75%) or high (25%) on the test. Overall, 12 (60%) of low portfolio writers achieved middle tests scores, while 4 (20%) had high test scores.

Effects of Gender

Mann-Whitney U-tests show no significant differences in test or portfolio performance for boys as compared to girls in any of the three grades.

Effects of Objectively-Measured Traits

MODE OF DISCOURSE

Distribution of modes of discourse differs from test to portfolio. Overall, narrative (41% vs. 46%) and exposition (19% vs. 16%) remain about constant in both test and portfolio samples. However, description (23% vs. 16%) and argument (15% vs. 6%) drop in representation, while poetry gains (2% vs. 16%). Whereas poetry shows marked increases in the portfolio in all grades, the drop in use of argument is restricted to grade 8 (23% to 8%) and grade 11 (28% to 11%), while remaining at 2% in both test and portfolio papers in fifth grade.

The gain in narratives chosen by fifth graders for the portfolios (62% vs. 49%) is matched by a decline in description (35% to 19%) and exposition (13% to 6%). In grade eight the shift in poetry from 3% of the test papers to 22% of the portfolio choices is matched by a drop in argument selections from 23% of the test to 8% of the portfolio. High schoolers actually included fewer narratives (23%

Table 10 Distribution and rankings of modes of discourse used in the test by grade level

TEST

Grade	Narrative	Description	Exposition	Argument	Poetry	Total	Rankings
5	55	40	15	2	1	113	A, D, N, E, P
(%)	(49)	(35)	(13)	(2)	(1)		
8	30	17	16	20	3	86	A, N, E, D, P *
(%)	(35)	(20)	(19)	(23)	(3)		
11	22	3	19	17	0	61	N, A, D, E **
(%)	(36)	(5)	(31)	(28)	(0)		
Total	107	60	50	39	4	260	
(%)	(41)	(23)	(19)	(15)	(2)		

* $p < .03$ ** $p < .05$

PORTFOLIO

Grade	Narrative	Description	Exposition	Argument	Poetry	Total	Rankings ***
5	206	62	19	7	36	330	
(%)	(62)	(19)	(6)	(2)	(11)		
8	102	38	38	20	57	255	
(%)	(40)	(15)	(15)	(8)	(22)		
11	38	23	64	18	25	168	
(%)	(23)	(14)	(38)	(11)	(15)		
Total	346	123	121	45	118	753	
(%)	(46)	(16)	(16)	(6)	(16)		

*** rankings not significant

vs. 36%) and arguments (11% vs. 28%) in their portfolios, but increased the representation of poetry (0% to 15%), description (5% to 14%) and exposition (31% to 38%).

Kruskal-Wallis tests show differences in test score among mode of discourse groups in grade 8 ($H(4) = 11.856$, $p < .03$) and grade 11 ($H(3) = 8.433$, $p < .05$), but not in grade 5. Table 10 (p. 60) shows the distribution of test papers by mode of discourse. In column 2 we can see that narratives account for half of the grade 5 test papers and at least one-third of the grades 8 and 11 papers. Column 5 demonstrates that arguments make up about one-quarter of test papers in the secondary grades but 2% of grade 5 tests. Column 3 shows a corresponding decrease in the number of descriptions from 40 of 113 in grade 5 to 17 of 86 in grade 8 and only 3 of 61 in grade 11.

Kruskal-Wallis tests of difference in paper score among mode of discourse groups in writing chosen for the portfolio show significance in only one of nine papers (3 papers per portfolio over 3 grades). In grade 5, one of three paper sets contains differences among types of writing significant at $p < .02$ ($H(4) = 12.08$).

RANGE OF MODES OF DISCOURSE

In the total sample ($r = -.133$, $z = -2.093$, $p < .03$) and in grades 5 ($r = -.196$, $z = -1.788$, $p < .02$) and 11 ($r = -.233$, $z = -1.788$, $p < .04$) range of modes of discourse in the portfolio inversely correlates with the length of the portfolio. Those who chose a wider array of pieces wrote shorter pieces in elementary and high school, and in general across the whole population. Also across the entire sample, the range of both the high test score group ($r = -.241$, $z = -2.284$, $p < .01$) and the low test group ($r = -.45$, $z = -1.857$, $p < .03$)

correlated inversely with length, but range of choices of the middle group did not.

In grade 8, range directly predicts portfolio score (r = .203, z = 1.882, p <.03). Only in grade 11 do other range scores correlate with writing scores. For the low test score eleventh graders portfolio score is negatively predicted by range (r = -.756, z = -2.00, p <.02). For high test score juniors, the range negatively predicts test score (r = -.443, z = -1.717, p <.05).

Socio-economic status also predicts range of works chosen for the portfolio. Across the whole population, range of the portfolios of the middle test score group directly correlates with SES (note: numbers are negative because the higher the SES number, the lower the actual SES: r = -.443, z = -1.717, p<.05). For the middle test group in grade 8, SES positively correlates with range (r = -.427, z = -2.526, p <.006). In grade 5 the SES of the middle test group (r=-.35, z = -2.948, p <.002) and the high test group (r = -.328, z = -1.734, p <.04) also correlates directly with range.

MONTH AND LENGTH OF WORK

Spearman tests show no correlation between mean length of work on papers in a portfolio and the median month the student began work on the portfolio pieces in any of the three grades. Only in grade 8 did month of work positively and significantly correlate to portfolio score (r = .208, z = 1.834, p < .034), while in grades 5 and 11, the relationships were insignificant, but negative in direction (gr. 5: r = -.125 ; gr. 11: r = -.06).

Table 11 (p. 63) displays the gain in portfolio score across the school year in grade 8. Portfolios whose median month of start is September or October achieve the lowest scores (columns 2 and 3). November and December starts (columns 4 and 5) are significantly

Table 11 Differences in grade 8 portfolio scores (PM) by median month of work

	Sept.	Oct.	Nov.*	Dec.**	Jan.	Feb.
PM	4.00	4.33	5.43	5.36	4.50	4.93
(SD)	(.58)	(.89)	(1.40)	(1.08)	(1.31)	(1.11)

* greater than Sept. ($p < .02$)

** greater than Sept.($p < .035$) and greater than Oct.($p < .038$)

higher than portfolio selections begun earlier, and approach significance over January and February (columns 6 and 7) starts.

Across the three grades Kruskal-Wallis tests found significant differences in month, length and duration of work. Table 12 (p. 65), column 2 indicates average start dates for portfolio papers of students in all three grades fall in December, the fourth month of the school year. Column 3 indicates large differences in average length of portfolio papers, with grade 8 papers actually being shorter than those in grade 5. In column 4 we can see duration of work on a piece increases from two weeks in grade 5 to nearly a month in grade 8, but falls dramatically to under a week in grade 11. A Spearman rank correlation test shows significant, direct correlations across the grades between portfolio score and length (r = .154, z = 2.364, p <.009) and portfolio score and duration (r = .418, z = 6.601, p <.00003), but not month of work.

SCHOOL ADMINISTRATIVE UNIT (SAU)

Kruskal-Wallis tests of difference in test and portfolio scores among SAUs indicate no significant differences exist, except in test score at grade 5 ($H(9)$ = 19.343, p <.025).

Pairwise comparisons of grade 5 test scores of SAU show that only three SAU significantly outperform any of their peers. In column 5 of Table 22 (p. 92), SAU A (5.7) rates higher than E (3.8) or H (4.1), SAU D (5.4) is higher than H or J (4.1), and SAU F (5.0) rates significantly higher than H.

SOCIO-ECONOMIC STATUS (SES)

Socio-economic status (SES) of the school from which a student comes directly affects a student's test and portfolio scores, except

Table 12 Mean month, length and duration of work for portfolio papers across grade levels

Grade	Month (of school year)	Length (in words)	Duration (in days)
5	4.5 (1.2)	368.5 (492.9)	14.8 (16.9)
8	4.2 (1.4)	295.6 (198.7)	28.5 (51.6)
11	4.8 (1.4)	501.9 (347.9)	5.9 (7.9)
significance of differences	p <.015	p <.001	p <.001

for portfolio scores in the high school sample, as shown in Table 13 (p. 67), column 3.

Since the higher the number used to rate SES, the lower the actual status, the tests produced negative rs. In addition, SES accounts for no more than 5.7 % of portfolio performance (Table 13, column 2, r = -.239). Spearman tests of SES and test score correlation within low, middle and high test score groups did produce one significant r, however. In Table 14 (p. 68) column 3, R^2 = .45, SES thus accounting for 45% percent of test performance for the low test group in grade 5. R^2 figures for the middle and high groups are only .01% and .2%, respectively.

Spearman tests of rank correlation in Table 15 (p. 69) demonstrate that SES predicts longer portfolio length and duration of work across the whole population (columns 3 and 4). In grade 11 SES correlates with duration of work (r= -.344, z = -2.618, p <.0045), and in grade 8 only length of portfolio significantly correlates with SES (r= -.299, z = -2.745, p <.0031).

Measures of Writer Judgment

STUDENT EXPECTED SCORES AND SELF-RATINGS

Using Spearman rank correlations, I find that fifth (r = .188) and eighth (r = .399) grade students (Table 16, p.70, columns 2 and 3) significantly predict the scores adult raters assign their portfolios, while eleventh graders (column 4) do not. Only the self-ratings (SR) of grade eight (r = .423) students (column 3) significantly correlate to adult ratings (PM). In all cases, student expected grades (SE) and self-ratings exceed the actual ratings received by at least 1 point on a seven-point scale.

Table 13 Correlation of socio-economic status (SES) with test score and portfolio score by grade level

	Grade 5	Grade 8	Grade 11
Test			
n	114	86	62
r	-.17	-.21	-.23
z	-1.83	-1.97	-1.76
p	<.03	<.025	<.039
Portfolio			
n	114	87	60
r	-.23	-.24	-.15
z	-2.48	-2.21	-1.11
p	<.007	<.014	n.s.

Table 14 Percent of variance of test score versus portfolio score accounted for (R2) by socio-economic status (SES) of high, middle and low test score groups in grades 5, 8 and 11

Group	(n)	Test R^2	(n)	Portfolio R^2
Grade 5				
High	(31)	.2	(30)	9.0
Middle	(74)	.01	(79)	1.0
Low	(9)	45.0	(5)	0
Grade 8				
High	(47)	.04	(23)	1.2
Middle	(36)	.4	(53)	1.5
Low	(3)	*	(11)	0
Grade 11				
High	(17)	8.0	(15)	1.7
Middle	(36)	.8	(41)	1.9
Low	(9)	1.1	(4)	*

* too few to calculate

Table 15 Correlation of socio-economic status (SES) with month, length and duration of work in the total population

	Month	Length	Duration
n	245	252	238
r	.06	-.13	-.19
z	1.01	-2.01	-2.84
p	n.s.	<.02	<.002

Table 16 Student expected scores (SE), student self-ratings (SR) an portfolio scores (PM) in grades 5, 8, and 11

	Grade 5	Grade 8	Grade 11
SE	6.15 *	5.94 **	6.05
(SD)	(1.35)	(1.32)	(.99)
(n)	(112)	(87)	60)
SR	6.17	5.92 ***	6.07
(SD)	(1.45)	(1.09)	(1.09)
(n)	(113)	(87)	(60)
PM	4.73	4.64	4.79
(SD)	(1.31)	(1.41)	(1.33)
(n)	(115)	(87)	(61)

* correlated to PM at p<.05 ** correlated to PM at p<.0001 *** correlated to PM at p<.0002

When the correlations are broken down by test score group, Spearman tests of correlation show that no low score group in any of the grades significantly correlated with teacher ratings in either expected grade (SE) or self-rating (SR). Among middle test score groups, only eighth graders significantly matched adult ratings, doing so in both SE (r = .424, z = 2.506, p <.006) and SR (r = .527, z = 3.117, p <.001). In grade 5 (r = -.348, z = -1.872, p <.03) and grade 8 (r = .405, z = 2.79, p <.0026), the high scoring test group significantly predicts teacher ratings with their expected scores (SE). Only in the case of high scoring eighth grade test takers do student self-ratings (SR) significantly correlate with actual portfolio scores (r = .286, z = 1.937, p <.027).

PAPER STRENGTHS

A four-way frequency analysis developed a hierarchical log-linear model of the effect of match between rater and writer judgment about the strengths of papers on the portfolios score (P). Polytomous variables included portfolio score group (high, middle, low) and positive, negative or no agreement on the conception (C), language (L), and emotion (E) of papers. A positive match on conception indicates the writer and the rater both chose either ideas or organization as a strength of one portfolio paper. If writer and rater positively match on language, both find the wording or mechanics to be strong. When both rater and writer tie the value of the writing to the flavor of the piece or the experience of the writer producing or sharing it, a match results on emotion. Negative matches indicate both parties omit the category as a strength. No agreement occurs when one includes the category, while the other omits it.

In grade five, 115 students; in grade eight, 87; and in eleventh grade, 61 students provided portfolios of up to three papers and lists of up to three strengths for each paper. Two raters read, graded and listed up to three strengths for each paper. Since many of the 81 cells in the multiway frequency table contained zeroes, too many marginal tables fitted to the model also contained zero cells. In line with Knoke and Burke (1980), I added .5 to every cell. Knoke and Burke call this "a conservative procedure which will tend to underestimate effect parameters and their significances" (p.64).

Stepwise selection by simple deletion of effects using BMDP4F produces models at the fifth and eleventh grade levels including only first-order effects (grade 5: E,P,L,C -- $G^2(46) = 40.35$, $p < .7071$; grade 11: L,E,C,P -- $G^2(46) = 39.17$, $p < .7518$). In grade eight the best model found includes one second-order association, that between the portfolio score and a match on emotion (L,C,PE: $G^2(32) = 20.63$, $p < .9394$). All models chosen meet both standards suggested by Tabachnick and Fidell (1989). First, it should be the model including the fewest effects with a non-significant likelihood ratio chi-square (G^2), indicating a good fit compared to the saturated model. Second, it should have a non-significant difference from the next most complicated model. A summary of the models with results of tests of significance (partial likelihood ratio chi-square, G^2) and loglinear parameter estimates in raw and standardized form appears in Table 17 (p.73). Effects within each grade level are listed in order of the absolute value of the standardized deviates (Lambda/SE). Absolute values larger than 1.96 indicate significance in a two-tailed test of $p < .05$.

Table 17 Partial association likelihood ratio X^2 (G^2) and standardized loglinear parameter estimates (Lambda/SE) for significant effects in grades 5, 8 and 11

Grade 5

Effect	G2	Lambda/SE		
		None	Positive	Negative
Language	63.24*	.541	6.900	-4.763
Conception	102.26*	-6.667	0	6.667
Emotion	102.26*	2.512	5.041	-3.934
		Low	Middle	High
Portfolio Group	45.97*	-4.719	5.833	1.181

Grade 8

First-order effects

		None	Positive	Negative
Emotion	105.34*	1.045	5.379	-3.390
Language	80.85*	1.863	4.933	-3.590
Conception	162.93*	-5.650	0	5.650
		Low	Middle	High
Portfolio Group	17.68**	-3.503	4.653	.427

Second-order effects:

Port. by Emotion	15.91***		Low	Middle	High
		None	3.633	-1.994	-1.220
		Pos	0	0	0
		Neg	-3.633	1.994	1.220

Grade 11

		None	Positive	Negative
Emotion	54.24*	1.561	4.465	-3.244
Language	13.32****	1.812	2.462	-3.153
Conception	104.50*	-5.136	0	5.136
		Low	Middle	High
Portfolio Group	26.79*	-3.330	4.608	0.362

* p .0001 ** P <.0001 *** p <.003 **** p <.001

At all grade levels significantly more middle score portfolios occur than expected: 68 of 115 (59%) in grade 5, 47 of 87 (54%) in grade 8, and 38 of 61 (62.3%) in grade 11. Also at all grade levels fewer than predicted low score group portfolios appear: 11 of 115 (9.6%) in grade 5, 16 of 87 (18.4%) in grade 8 and 6 of 61 (9.8%) in grade 11.

In grades 5, 8 and 11, virtually all raters and writers agree that conception is a strength of the portfolio papers: 98 of 115 (85.2%) fifth grade portfolios, 84 of 87 (96.6%) of eighth grade ones, and 57 of 61 (93.4%) in eleventh grade. In grade five, more students than predicted, 17 of 115 (14.8%), showed no agreement with raters on the conceptual strength of the portfolio papers. In grades 8 and 11 only 3 and 4 portfolios, respectively, showed no agreement with raters on conception. No portfolios at any grade level demonstrated negative agreement on conception.

More than expected portfolios in all the grades showed agreement on the strength of language. Of fifth grade portfolios, 76 of 115 (66.1%) positively agreed that language was strong, while fewer than expected, 9 (7.8%), showed negative agreement. In grade 8, 63 of 87 (72.4%) positively agreed, while only 1 (1.1%) negatively agreed. At the high school level, 28 of 61 (45.9%) both listed language as a strength, and 8 (13.1%) both left it off the list.

Again in all grades, more students than expected and their raters both listed the emotional aspects of the writing as a strength, while fewer than expected left emotion off the list of strengths. In grade five, 80 of 115 (69.5%) show positive agreement, 1 (0.7%) negative, and 34 (34.8%) no agreement at all. In the middle school sample, 80 of 87 (92%) positively agree and 1 (1.1%) negatively match. In grade eleven, 44 of 61 (72.1%) both

list emotion as a strength, but 1 (1.6%) portfolio occurs where both raters and writer fail to mention emotion.

Finally, Table 17 indicates significant cell deviates for the mention of emotion by eighth graders in the low and middle score groups. Of 16 in the low group 8 (50%) disagree with raters as to whether or not to list emotion, but none agree with raters that it should be left off. Of 47 middle score papers, 3 (6.4%) show no agreement and 1 (2.1%) shows negative agreement.

Table 18 (p. 76) shows that in grades 5 and 8 the most positive agreement appears in the high score row, whereas in grade 11, the lowest ratios appear in the high score row. Moreover, examining column 4 (Conception), we can see that within each grade approximately equal proportions of students in low, middle, or high groups match adult raters on the importance of information and organization in the portfolio papers. Meanwhile, higher proportions of students in the higher score groups match raters on the strength of the flavor or the writer's experience (column 2).

Table 18 Percentage of writers by portfolio score group matching rater judgment of strength of paper (Emotion, Language, Conception)

Grade 5

Group	Emotion	Language	Conception
Low	54.5%	63.6%	81.8%
Middle	72.0	58.8	83.8
High	69.4	80.5	88.8

Grade 8

Group	Emotion	Language	Conception
Low	50.0	81.2	75.0
Middle	93.5	100	69.6
High	91.3	100	78.3

Grade 11

Group	Emotion	Language	Conception
Low	83.3	50.0	100
Middle	76.3	44.7	97.4
High	58.8	47.1	82.4

Chapter 5

Discussion

Tests Ineffective, Biased Predictors of Classroom Writing

Prompted-essay exams and evaluations of portfolios of classroom writing produce essentially the same rankings of students' written products, since scores significantly correlate at all grade levels (Table 7, p.56). Despite significant correlations, low R^2 figures (Table 7) and Cramer's V ratings (Table 8, p.57) indicate tests do a poor job of predicting future classroom writing. Specifically, observed frequency tables (Table 9, p.58) indicate that 80 of 114 (70%) fifth graders fall into exactly the same test and portfolio score groups, 44 of 86 (51%) in the total sample.

But the failure of tests to predict classroom performance, as measured by the portfolios, does not impact all score groups equally. In the total population, 20 students achieved low test scores and might be expected to be excluded from programs or retained in others, if the test were used to screen for ability. However, 16 (80%) of those so excluded actually scored average or above ratings on their classroom work. In contrast, only 4 (4.26%) of those achieving high test scores, and 13 (8.97%) of the middle test score group achieved low portfolio ratings, and might, therefore, have been erroneously admitted or promoted to a

program. Still, 16 of 20 (80%) low portfolio scorers achieved middle or high test scores, and were, therefore, missed by the screening test.

Clearly, prompted-essay tests of writing ability do a poor job of identifying potential classroom writing failures, while excluding many who will actually perform better in class. Moreover, tests will erroneously exclude far higher percentages than they will mistakenly admit.

Tests Least Accurate in Grade 8

Table 9 (p. 58) indicates that 51% (44 of 86) of grade 8 students fall into exactly the same test and portfolio groups. That is, those who score below, at or above average on one measure score in the same group on the other. In grades 5 and 11 the percentages are 70% and 65%, respectively. Figure 2 (p.53) shows that no eighth grader scored 2 on the test, but three did in portfolio writing, while overall 11 eighth graders received low portfolio scores, compared to 3 low test scorers.

The test score advantage in grade 8 may result from the nature of the response dictated by the test situation and the particular prompt used. Table 20 (p.80) reports the qualities grade 8 raters found in the papers they scored. In it, under "Overall Qualities," they

Table 19 Qualities of grade 5 papers by rater score as generated by raters April 11, 1990

1-Papers	
Strengths	***Weaknesses***
good initial ideas	ideas not developed
	confusion
	no focus
	poor mechanics
	tend to be short
2-Papers	
Strengths	***Weaknesses***
more developed than ones	bed-to-bed stories
have a beginning, middle and end	no feelings or passion
	an account of an event
	"impossible fiction"
	undeveloped ideas
	violence a substitute for plot
	no originality
	no narrative mixed with dialogue
3-Papers	
Strengths	***Weaknesses***
organization	no attempt to do anything with
teacher corrections	
original plot	weakness in the plot
humor	poor endings
better mechanics	no effort to polish form
	lack strong personal voice
	not as much investment in topic
4-Papers	
Strengths	***Weaknesses***
all the basics plus: a spark	mechanics
natural flow	loose focus
very fluent	vocabulary
sense of the genre	
switch back and forth	
between modes	

Table 20 Qualities of grade 8 papers by rater score as generated by raters April 12, 1990

1-Papers	
Strengths	***Weaknesses***
legible	lack of interest
breaking into writing	lack of detail
	lack of effort
2-Papers	
Strengths	***Weaknesses***
had high points, moments	no development of beginning, middle, end
maybe just one line saved paper	lack of organization, didn't know where to go
had at least a beginning, middle, end structure	interest to reader
	undistinguished
3-Papers	
Strengths	***Weaknesses***
interesting to reader	just not perfect
imagination	more dynamics
vocabulary, analogies, metaphor	details
learned phrases ("Founded on the premise...")	more varied sentence structure
almost perfect mechanics	
more than just getting JOB done	
interest, feeling, purpose	
focus stronger	
something special said	
4-Papers	
Strengths	***Weaknesses***
Original, Wonderful humor, focus	
student's/teacher's criteria matched	NONE

Overall Qualities

1) Prompt helped on the test
2) Lack prompt/purpose confusing with portfolio papers

Table 21 Qualities of grade 11 papers by rater score as generated by raters April 13, 1990

1-Papers	
Strengths	***Weaknesses***
Risk-taking	Generalization
Emotional ties	Mechanics
	Lack of focus
2-Papers	
Strengths	***Weaknesses***
Excitement / power behind ideas	Inconsistencies in focus, effect and use of language
Better developed sense of language	Lack of honest tone
3-Papers	
Strengths	***Weaknesses***
Organization	Lack of maturity
Mechanics	Poor mechanics
Development and support of ideas	
4-Papers	
Strengths	***Weaknesses***
Sincerity	
Flavor	
Strong mechanics	none
Exploration of idea with all its ramifications	

Problem Areas in Writing:

1) Lack of interesting vocabulary
2) Lack of personalization

note,"Prompt helped on the test." Compared to the more specific and scripted prompts of earlier years (e.g., "If you could have dinner with any three people, living or dead, who would they be and why?"), the general prompt used ("Write about something you know about and care about. Make sure your reader knows how much you know and how much you care.") produced writing with more personal commitment, several raters said.

But consider the distribution of modes of discourse in the test as opposed to the portfolio selections (Table 10, p.60). Poetry comprises only three per cent of test papers but 22 per cent of portfolio selections in grade 8. Conversely, arguments account for 23 per cent of test papers but only eight per cent of portfolio pieces. Since argument rates first in grade 8 test scores and poetry last, students' decisions not to write poetry for the test clearly accounts for the difference in scores.

Consider that eighth grade writers (Table 12, p.65) wrote the shortest pieces (296 words, versus 369 in grade 5 and 502 in grade 11), but that they required the longest duration of work (29 days versus 15 in grade 5, or 6 in grade 11). Since virtually no one at any grade level wrote poetry for the test (four of 260 papers, Table 10, p.60), the test setting seems to militate against the choice of poetry. But the availability of choice in portfolio selection may have affected the scores of eighth grade writers, since they chose more poetry than their younger or older peers (22 per cent versus 11 per cent of fifth grade papers and 15 per cent of grade 11 pieces, Table 10, p.60). This finding supports Newkirk (1984) who

predicts that writing for peers, which classroom writing may often be, may fail academic requirements.

Results Consistent with Theory and Other Studies

These results confirm findings by Buddmeier and Raivetz (1990) and McLarty et al.(1985) that test results do not correlate well with other, more context-bound measures of performance. The results duplicate the findings with 28 fifth-graders in Simmons (1990) and confirm the theories of White (1985), Moffett (1981), Elbow and Belanoff (1986a) and Gardner and Hatch (1989), which claim that ratings based on several samples selected by the writers and developed over a period of time in naturalistic settings describe writing abilities better than one-shot, scripted essays.

The children quoted by Comstock (1988) claim writing taken from their folders will show how well they can write. Martin (1988) also finds extemporaneous essays to be superfluous to the pass/fail decision, when portfolios of pieces developed during the course are available. The current results and these theories contradict Anastasi (1982), who claims that controlled tasks and settings are necessary to produce valid, reliable measures of writing ability.

Essay exams serve a gate-keeping function by granting or denying people access to programs based on their predicted ability to perform writing tasks in those future contexts. But 80% of those who scored low on the test in this sample eventually

produced average or above average ratings on their classroom pieces, while 80% of those who eventually "failed" (by scoring low in the portfolio) had achieved middle or high test scores. The current results strongly support earlier injunctions that we abandon such isolated testing as a gate-keeping device.

Our society often refuses to promote, graduate, employ or fund based on test results; therefore, we must ferret out and eradicate testing bias against any one group. Table 9 (p.58) clearly indicates that isolated tests of writing ability, when used to screen applicants for admission to future programs or graduation from present ones, will both exclude many who can perform and admit others who cannot. However, since 80% (16 of 20) of the excluded low test group performed adequately, while only 6.7% (16 of 240) of admitted middle and high test scorers failed in the portfolio measure, tests clearly hurt the worst performers more than they help their more apt peers.

Not only do tests hurt the worst test-takers, they also hurt the poorest schools. Table 14 (p.68) indicates socio-economic status of the school predicts 45% of test score for the low test score group in grade 5, versus negligible amounts for the average and above average groups. Too few low scores occurred in the middle school sample for conclusions to be drawn. In the high school sample, where students worked a very short time on portfolio assignments, test and portfolio scores did not differ significantly.

Thus, in the only adequate sample where the conditions for daily writing differ markedly from test conditions, students from poorer schools can generally be expected to perform poorly on tests.

Meanwhile, their peers from wealthier areas do not suffer in test situations. Finally, no similar connection exists for these fifth-graders between SES and portfolio score.

States such as Virginia, Maine, and Vermont where individuals or districts are to be evaluated on the basis of extemporaneous, prompted-essay tests must heed these results. Any child from a poorer school who is excluded from one program, or forced into another, based on an essay test would seem to have a legitimate basis to claim unfair discrimination. Poorer school districts, or their leaders or teachers, that suffer public embarrassment, or loss of professional advantage, from open reports of deficit based on such tests should also complain.

Effect of Gender

Probably due to the size of the samples (n=115, 87, 61), the gender effect found in Simmons (1990) with only 28 subjects fails to appear in this study. This finding supports the position of Hyde and Linn (1989) that sex differences in verbal ability have been erased. If males and females conceive of story differently, as predicted by Gilligan (1982), their raters seem able to find value equally in male or female conceptions of story.

Mode of Discourse Effects

Table 10 (p.60) shows students clearly prefer to write narratives, even in a timed, test situation, since narratives dominate the selections of all grade levels in the test, and in the

portfolios of grades 5 and 8. However, since previously marked coursework constituted the majority of eleventh grade portfolio papers, it would seem high schoolers may have included more exposition in their portfolios due as much to teacher assignment as to personal preference. High school raters complained that the portfolio pieces lacked "personalization" (Table 21, p.81) and actually preferred reading the test papers. Ironically, narratives (36%) outnumber exposition (28%) in the eleventh grade test, and the raters themselves, as high school teachers, gave the assignments that produced such deadly portfolio writing. It would seem high school teachers value one set of qualities when they read student papers as samples of writing ability, but use another set when they construct writing assignments to transact their daily classroom business. These results confirm predictions by Freedman (1983) that forced-choice tasks may produce duller writing.

Table 10 (p.60) clearly indicates that students choose very different writing tasks to represent their abilities when provided with choice, time and a supportive community. Whereas poems account for only four of 260 (2%) test papers, written in isolation over 90 minutes, 118 of 753 (16%) portfolio selections are poems chosen from pieces written in the classroom during the course of the year. And, the writers seem to have exercised good judgment by limiting their poetry writing to the portfolio. Although poems ranked lowest in the ratings in grade 8 and 11 test papers (Table 10, p.60), they did not differ significantly in score from other

modes of discourse in portfolio rankings.

Since grade 8 shows the biggest increase in poetry in the portfolio, and eighth graders matched adult judgment in both paper score (Table 16, p.70) and strength of paper (Table 18, p.76), their more frequent inclusion of poetry seems based on mature awareness of both the writing context and the values of their readers.

Growth Marked by Choice of Mode

Distribution of narratives relative to exposition and argument reflected in Table 10 (p.60) also supports Moffett's (1981) definition of growth as doing more things and the old things better. Narrative, a form learned earlier in school, declines as a percentage of the sample in both the test (Table 10: gr. 5, 49%; gr.8, 35%; gr.11, 36%) and the portfolio (gr.5, 62%; gr.8, 40%; gr.11, 23%) from grade 5 through high school, but the effect is more pronounced in the portfolio selection. At the same time, argument and exposition, later emerging forms, consistently increase from a total of 15% of grade 5 tests, to 42% in grade 8, and 59% of eleventh grade tests. Likewise, argument and exposition grow from 8% of grade 5 portfolio pieces, to 23% of those selected in grade 8, to 49% in grade 11.

Based on Crowhurst and Piche (1979). Lunsford (1981), and Flower (1981) these figures indicate increasing ability to use less narrative and more argument and exposition when time demands it (on the test), or to select (for the portfolio) the forms requiring more abstract thinking, greater distance between the writer and the audience, and more cognitively developed syntactic structures. Therefore, student choice, either in the form of more general

prompts for essay tests or control of portfolio construction, provides a measure of growth over time abrogated by systems that require pieces from certain genres, as in Vermont, or ones that give selection to the teacher. The findings also support Gardner's contention that measures of student reflection and judgment are required to demonstrate development of ability.

Method of collection may explain the failure of range of modes of discourse included in the portfolio to correlate as consistently with portfolio score as in Simmons (1990). For that study, a single rater sorted all papers into mode of discourse groups. For the current study more than a thousand papers arrived from many classrooms and schools spread across eleven school administrative units. Due to the short time available between arrival, rating and return to the student, the individual classroom teachers marked papers for mode of discourse group, obviously increasing the possibility of error in this measure.

However, significant correlations do occur with paper scores in grades 8 and 11. Notably, in eighth grade the inclusion of more types of writing predicts a higher portfolio score. Therefore, those who write and value a range of modes of discourse demonstrate the ability to produce more highly rated classroom writing. In grade 11 the trend reverses itself. There, those who score poorly on the test but choose a variety of pieces for their portfolios, actually achieve worse scores. Those who score well on the test actually choose fewer modes to represent their talent in the portfolios.

This finding in part corroborates the results of Simmons (1990) where higher scores generally reflected the ability to use and value more types of writing. These results also support Moffett

(1981) and can be used to argue against requiring certain types of writing in a portfolio, since the requirement would rob students of the chance to demonstrate this predictive power of choice. The appearance of the predictive power in grade eight may indicate growth over grade 5, but the reversal of the trend in the high school sample would, therefore, indicate an erosion of that ability by grade 11.

Range most strongly correlates with socio-economic status in the current study. For the middle and high test score groups, higher SES predicts a wider range of choices in the portfolio. Since SES also correlates with duration of work (see next section), it may be that schools from wealthier towns are providing students with more chances to keep their work open longer and to experiment with a variety of types of writing.

Month and Length of Work

Students at all grade levels pick their best work from pieces done slightly closer to the end of the school year than the beginning. Students assembled portfolios during the last two weeks of the seventh month (March), and the mean month of selection for all grades was December. In addition, in grade 8, where paper score correlates with month of work, December papers outranked September starts, but January and February pieces dropped in score after the long vacation (Table 11, p.63). Students, therefore, tend to make progress during the school year and recognize that progress. Significantly, the grade 8 sample also produced the highest correlations of teacher/student judgment (Table 16, p.70, and Table 18, p.76) and the highest scores on the

prompted test taken in March. Thus, the growth in ability reflected in the dates of portfolio work is supported by the degree to which the writers match adult judgment and perform on de-contextualized measures of ability.

These results confirm the positions in White (1985) Moffett (1981) Gardner and Hatch (1989) and Greenberg and Witte (1988) that periodic sampling of student work will indicate growth, and therefore, that more than one sample of writing will be needed to reflect the development of writing ability. Moreover, these results indicate that both multiple samples and a mixture of both subjective ratings and objective measures can more fully evaluate writing programs than mere test results (Wolfe, 1990; Flood & Lapp,1989; Sevigny,1981).

High School Program Questioned

Table 12 (p.65) provides perhaps the most powerful program evaluation data in this study. Since across all age levels length ($p < .009$) and duration ($p < .00003$) correlate with portfolio score, it is striking to note that high school students, while writing the longest pieces, utilize the shortest duration of work. Moreover, since duration doubles between fifth grade (14.8 days) to eighth grade (28.5 days), we can only be more shocked by the drop to less than a week (5.9 days) for the oldest, and presumably the most cognitively developed, students.

High school scorers (typically English teachers) complained that so many history papers and lab reports were included, and the training sample proved to be mostly classroom assignments. Clearly, high school writing tends to be done in response to teacher assignment, generally on a short deadline. Further, high school

students, as opposed to their middle and elementary school peers, seldom re-open tasks for further work. Graded work in high school is truly dead in the water.

These findings are particularly troubling in light of the Getzels and Csikszentmihalyi (1976) finding that keeping the problem open to revision longer produces more creative work. It would appear that high school teachers, in order to use writing as demonstration of knowledge of course content, not only fail to address the development of writing abilities, but actually create a climate aversive to their continued growth.

As will be seen below in the discussion of measures of teacher and student judgment, this climate for growth in writing includes not only control of subject, form, and duration of work, but understanding of each other's standards and expectations as well. When we realize that high school students, when given the choice by a general testing prompt, choose to write proportionally more narratives, even in a shorter testing situation, than they are able to select from their portfolios, we can see who controls their daily writing choices. When we note that they seldom keep a paper open longer than during the week in which it is assigned, we know who decides when work should stop. When the high school raters, themselves teachers of these students, prefer the personal flavor of the tests to the lack of commitment in writing they themselves assign (Table 21, p.81), we can only conclude students probably receive mixed messages about writing and will have trouble matching or anticipating teacher judgment.

Few Differences Among SAU

Scores from the twelve school administrative units differed

Table 22 Group means of grade 5 papers by score group and by SAU

Grade 5 Profiles

SCORE GROUPS

GROUP	SES	n	TEST	PORT	RANGE	MONTH	LONG	DURAT	CONC	LANG	EMOT
ALL	393	114	4.8	4.8	1.8	4.5	360	14.8	S	S	S
HIGH	304	31	6.6	5.7	1.8	4.6	430	15.5	89%	81%	69%
MID	421	74	4.3	4.6	1.8	5.1	292	14.2	84	59	72
LOW	474	9	2.7	4.1	1.8	2.4	879	17.4	82	64	55

S = significant

SAU GROUPS

SAU	SES	SES GR	n	TEST	PORT	RANGE	MONTH	LENGTH	DURATION
A	235	HIGH	10	5.7*	5.1	2.0	6.4	246	6.5
B	146	HIGH	5	4.2	4.2	1.6	5.2	311	5.3
C	110	HIGH	33	5.2	5.2	1.9	4.2	596	27.6
D	379	HIGH	11	5.4**	4.6	2.1	4.5	205	5.2
E	462	LOW	3	3.8	5.2	1.3	5.3	407	24.7
F	483	LOW	8	5.0	5.0	1.7	6	259	26.3
G	055	HIGH	10	4.7***	5.3	2.0	4	224	5.5
H	618	LOW	14	4.1	4.1	1.7	4.4	271	7.6
I	1240	LOW	10	4.6	4.5	1.4	5.4	372	9.0
J	708	LOW	6	4.1	4.6	1.1	4.8	240	14.2

* A> E, H $p < .05$ ** D > H, J $p > .05$ *** F > H $p < .05$

Table 23 Group means of grade 8 papers by test score group and by SAU

Grade 8 Profiles

Test Score Groups

GROUP	SES	n	TEST	PORT	RANGE	MONTH	LONG	DURAT	CONC	LANG	EMOT
ALL	351	86	5.6	4.7	1.85	4.4	296	29	S	NS	S
HIGH	310	47	6.5	5.1	1.9	4.6	322	26	78%	100%	91%
MID	374	36	4.5	4.4	1.7	4.0	274	33	70	100	94
LOW	810	3	3	3.3	2	4.3	132	5	75	81	50

s = significant ns = non-significant

SAU GROUPS

SAU	SES GRP.	#	TEST	PORT	RANGE	MONTH	LENGTH	DURATION
A	H	12	6.2	5.3	1.6	4.8	451	10
B	H	6	5.2	4.7	2.2	6.8	206	31
C	H	20	5.7	5	2.3	4.1	339	31
D	H	9	5.3	4.3	1.7	4	192	2
E	L	5	6	5.2	2	4	421	9
F	L	5	5.4	4	1.2	4.5	198	4
G	H	6	6.5	5	1.7	4.3	390	17
H	L	12	5	4.8	1.6	4.3	288	96
I	L	3	4.7	3.7	2	4.5	89	2
J	L	8	5.1	4.3	1.9	3.1	140	11

Table 24 Group means of grade 11 papers by test score group and SAU

Grade 11 Profiles

TEST SCORE GROUPS

GROUP	SES	n	TEST	PORT	RANGE	MONTH	LONG	DURA	CONC	LANG	EMOT
ALL	483	60	4.7	4.8	1.9	5	502	6	S	NS	S
HIGH	305	16	6.5	5.6	1.9	5.1	639	10	82%	47%	59%
MID	572	36	4.3	4.5	1.9	4.8	488	4	97	45	76
LOW	434	8	2.8	4.3	2	5.6	292	6	100	50	83

S = SIGNIFICANT NS = NOT SIGNIFICANT

SAU GROUPS

SAU	SES	n	TEST	PORT	RANGE	MONTH	LENGTH	DURATION
A	H	15	4.5	5.1	1.9	4.7	634	8.6
C	H	13	5.4	4.8	1.9	4.9	501	9.8
D	H	7	4.7	5.3	2.1	5	493	4.1
F	L	7	5.1	4.1	1.7	4.6	679	3.0
I	L	10	4.3	4.4	2.1	5.3	308	2.8
J	L	5	4	4.4	1.4	6	331	3.7
E	L	4	4.5	5	2	4.7	408	1.7

significantly only in grade 5 test scores, the group with the largest gap from low group mean (2.67) to high group mean (6.65). Since no other grade level demonstrated significant differences in test score among SAU, and since the measure of actual classroom performance, the portfolio score, produced no significant differences at any grade level, this finding seems to hold little importance.

Only one of the three significant pairwise differences in SAU test score (Table 22, p. 92) crosses socio-economic lines, however. SAU F of the low SES group (SES = 482) significantly outscores SAU H (SES = 618) also of the low SES group. Further examination of Table 22 shows that students from SAU H choose papers from the average month (4.4) for their grade level, while those from SAU F choose later work from February (month = 6). More striking, however, is the difference in duration of work. SAU H students work about half as long (duration = 7.6 days) as other fifth graders (overall average = 14.8 days), whereas writers from SAU F keep their writing open about twice as long as average (26.3 days).

Application of Results to Program Evaluation

These comparisons suggest that students from SAU H may be less aware of growth over the course of the year, while those from SAU F are more aware, although month of work did not correlate with score across the population. Since duration of work did correlate with score, the below average figure of SAU H may indicate a need for students to be encouraged to go back and review work for possible revision, or for the system to ask teachers if they determine when a piece is finished, rather than leaving that

choice for the writer. The much longer than average duration of SAU F students nearly matches that of eighth graders. Since duration also correlates with SES, it would seem SAU F may have been able to offset, in part, the effects associated with SES by providing a program in which students have more incentive to keep their work open longer. A portfolio evaluation of writing can lead us, therefore, more quickly to questions about the writing process in our classrooms, than to the traditional consideration of the characteristics of text.

But to focus on these questions would be to miss the forest while examining the trees. Over the past decade these same 12 SAU have routinely collected prompted writing samples, holistically scored them, and reported the mean score for each SAU in a formal report, complete with graphs of performance across SAU and years. Occasionally, superintendents have publicly cited these results as evidence of good or poor performance in their districts, and teachers have borne the brunt of the attack on inadequate programs.

Never, however, have holistic scores been viewed as the ordinal data that they are, nor have non-parametric inferential statistics been applied to them to determine where the significant differences actually lie. In fact, the test means listed in Tables 22-24 (pp.92-94) look a great deal like earlier reports, but few, if any, comparisons denote real difference. Further, since no significant differences appear in comparisons of actual classroom writing performance (the portfolios), the test score results must be taken with a large grain of salt.

Sadly, the superintendents who received these results seem to have taken another view of the matter. The testing program has

been under fire for several years, in part due to the public abuse of results mentioned above, and in part due to the general irrelevance of the holistic results and the intrusiveness of the isolated testing experience. Teachers widely have desired the program to go away. Faced with a horse race with no winners or losers and a chorus of disgruntled staff members, the superintendents have suspended testing.

If the Business Council for Effective Literacy (1990) is right, an increase in testing in adult literacy may indicate that field has come of age. Certainly, Glickman (1990) indicates teachers and schools who expect freedom and support in the nineties must be willing to demonstrate results, probably by some form of testing or assessment. Thus, teachers may err badly in merely resisting current testing without offering a substitute. Eisner (1990, November) says, "You can't beat something with nothing." In fact, the state of New Hampshire has recently moved to halt California Achievement Testing, but has rejected portfolio plans as too expensive, and is still investigating across state comparisons of performance (NHATE, 1990). In writing that would mean holistically or analytically scored, prompted writing samples -- same horse race, different track.

Although portfolio assessment of performance as evaluation of ability is in its infancy, the results of this study should show two things. First, a horse race analysis of isolated test essays will not capture actual writing ability as applied in natural contexts. Second, analysis of subjective and objective measures of student-selected portfolios will reflect the shape of programs that produce the writing and affect the writers.

Effect of Socio-economic Status (SES)

As argued above SES weakly predicts both test and portfolio score, except in the high school portfolio ratings, when all score groups are lumped together. SES does not predict equally across score groups, however. SES in grade 5 predicts 45% of the low test group's test score, but only negligible amounts of the middle and high groups' scores. SES does not have the same power in grade 5 portfolios.

In grade 8 only three low test scores occurred out of 86 tests; therefore, no correlation was calculated. However, examination of Tables 22, 23 and 24 (pp.92-94) indicates that the SES of the low score group in grade 8 lies dramatically below that of either the high and middle groups in grade eight (310, 374), or any of the groups in any grade (810 vs. a range of 304-572). Therefore, SES seems to retain its power to predict low test score in grade 8 as well.

In grade 11 a different picture emerges. First, Table 24 (p. 94) demonstrates that the SES of the middle score group actually is lower (572, the rating number, therefore is higher) than the low score group (434). Second, the average duration of work for the middle test score group is shorter than that of the low group (4 days as opposed to 6). Third, SES significantly correlated with duration of work in grade 11. Thus, the testing context -- a writing session limited to one day -- is less different from their normal writing behavior for high schoolers in general, and most markedly from that of the lowest SES group. By contrast, in grade 5 the low test group duration (Table 22, p.92) is the longest of the three groups (17.4), actually surpassing the high score group's 15.5.

At the high school level, then, we may have erased the effect of SES on test-taking ability by constructing all writing tasks as if they were tests. Ironically, of course, the high school raters preferred the flavor of test pieces written to a general prompt over the more rigidly assigned classroom essays included in the portfolios.

Measure of Writer Judgment

Student Expected Scores and Self-Ratings

Correlations of student expected scores (SE) and actual adult ratings (PM) reflected in Table 16 (p.70) confirm earlier findings that differentiate writing programs in the elementary and middle schools from those in area high schools. Students in the earlier grades significantly predict the holistic reactions of adult audiences, while their older peers do not.

This finding takes on more significance in light of the fact that the expectations of higher ability groups in grades 5 and 8 predict rater judgment, while those of low score groups do not. Also, eighth graders seem to have grown in their ability to reflect adult judgment over their fifth grade peers, since the expectations of the grade 8 middle test score group match their adult raters, as do the self-ratings of the middle and high scoring test-takers in the middle school. This apparent developmental trend is erased in the high school sample, where the ratings of no groups predict teacher judgment.

These findings confirm the theories of Gardner and Wolf that measures of student judgment and self-awareness will reflect the development of ability (Brandt,1988,;Hatch and Gardner,1986; Gardner and Hatch,1989; Wolf,1989). They also support the calls of

Flood and Lapp (1989), Della-Piana et al. (1988), Cooper (1990), and Bishop (1987,1989) for multiple measures to include the awareness of audience and the match between student and teacher judgment.

The findings demand serious attention to program constraints at the high school level, especially in light of the Getzels and Csikszentmihalyi (1976) finding relative to creativity and duration of work. Since the duration of work is shortest at the high school and no eleventh graders of any ability group are able to predict teacher rating, it seems they have few chances to receive feedback and re-open the task later for a meaningful re-write.

But the statements of high school raters also add evidence of mixed messages that might be confusing high school writers. Teachers who rated the test essays and portfolio pieces expressed a preference for the personal flavor of the essays, yet complained that they needed to know the exact teacher assignment in order to fairly judge the classroom writing. Clearly, they expected classroom assignments to be much more constraining than the test prompt. The high inter-rater reliability (.97 estimated by Cohen's Kappa) indicates they did not need to know the original assignment. This finding supports Martin (1988), who says she has declined as unnecessary her staff's requests for knowledge of the original assignments in their portfolio evaluation process.

These findings in conjunction lead to two conclusions. First, high school teachers expect to give classroom writing assignments that limit the writer's ability to express personal commitment and flair. Second, these same teachers use different standards to judge "good writing," than they use to construct and rate classroom writing exercises. Little wonder that eleventh grade writers could

not predict high school teachers' reactions!

Effect of Agreement on Strengths of Papers

Emotion

BMDP4F provides a loglinear analysis of effects of designated variables but does not make one variable dependent, as in logit analysis. In order to perform logit analysis with BMDP4F, we must find significant two-way associations. In grade eight, the second-order effect of emotion on portfolio score achieves significance, indicating that the best predictor of portfolio score is the degree to which student writers and adult raters both cite the flavor of the piece or the writer's experience as a strength of the work.

Significant standardized cell deviates occur for the low portfolio/no agreement and low portfolio/negative agreement cells, but no negative agreements occur in the low portfolios. Since Dixon (1988) says that sparse tables exaggerate the effect of low observed frequencies, we can safely ignore the effect of negative agreements. The same principle applies to the middle portfolio interactions with both no and negative agreement. Therefore, in grade 8 lack of agreement between teachers and students on the emotional strengths of the writing predicts low scores. Stated more positively, average and above portfolio writers tend to agree with their raters on the emotional strengths of the writing.

All first-order effects prove to be significant, but cell deviates again indicate the direction of the effects. Low observed frequencies for negative agreements can be ignored, but larger than predicted positive agreement occurs at all grade levels, highest in

grade 8 (92%) and about equal in grades 5 (69.5%) and 11 (72.1%). Saliently, 34.8% of grade 5 portfolios show lack of agreement on emotional factors -- the only grade where significant non-agreement occurs in this area.

Combined with earlier findings on the student expected scores and self-ratings, these figures indicate growth in maturity of judgment between grade 5 and grade 8, since more non-agreement on emotion occurs in the earlier grade, but the lack of agreement is not localized among low score portfolios. By grade 8 such disagreement has dropped in magnitude and does predict a low score. In the high school sample the percentage of positive agreement has dropped again, but the non-agreement is no longer localized to low score writers. This finding indicates a return to the grade 5 pattern of judgment, except that the percentage of responses for non-agreement (26.3%) is not significantly larger than predicted.

Language and Conception

Findings relative to matches on language and conception echo this pattern of progress from grade 5 to 8, followed by regression in grade 11. More than predicted fifth (66.1%) and eighth (72.4%) graders positively agree with raters about wording and mechanics being strong in the portfolios, while only 45.9% of high school readers and writers agree. Meanwhile, the percentage of raters and writers failing to find language a strength drops from 9% in grade 5 to 1.6% in grade 8, then rises to 13.1% in grade 11.

Although many positive matches on conception occur at all three grades, at no grade level do they exceed expectations. Low

numbers of non-agreement in grades 8 and 11 and negative agreement in all grades probably make them insignificant in a sparse table. But 14.8% of fifth graders fail to agree with adult readers on conception, an effect erased by eighth grade.

Portfolio Score

Differences between portfolio score group expected and observed frequencies, i.e. , more middle scores but fewer low scores than predicted at all grade levels, are probably trivial. Performance measures should group scores around the mean, not distribute them by chance across all three categories. Since observed high scores (32% in grade 5, 28% in grade 8, and 27.9% in grade 11) do not fall significantly under expectation, classroom performance seems to be better than predicted, despite the tendency of portfolio scores to lag behind test scores for the high group. It is also possible that the requirement for informed assent and consent from students and parents militated against low scoring students participating. However, the percentage of high-ranking portfolios remains relatively constant across grades, while participation rates (grade 5: 76%, grade 8: 71, grade 11: 46%) do not.

Agreement by Score Group

These results reinforce the pattern found in Table 18 (p. 76), namely that growth in writer judgment occurs from grade 5 to 8, but that eleventh graders fail to match adult judgment as well as their younger peers. Column 4 of Table 18 indicates that agreement on ideas and organization (conception) in grades 5 and 8 remains high and stable across score groups, but that high school high scoring writers actually agree less with adults than their lower scoring peers.

In column 3 low scoring eighth graders agree with their readers about as often as the high scoring fifth graders, with perfect agreement among the middle and upper eighth grade score groups and their raters. No high school group matches adult judgment as well as any of the fifth graders. Moreover, the writers of the highest rated grade 11 portfolios again do slightly worse than their lower scoring peers.

Differentiation of agreement scores in column 2 explains the effectiveness of likeness of judgment on emotional factors as a predictor of score. In both grade 5 and grade 8, many fewer low rating portfolios exhibit agreement with adults on the flavor or writer's experience. In grade 8, nearly twice as much agreement occurs in the middle and high groups as in the low group. On this significant factor, agreement reverses among high schoolers, the lowest scoring group agreeing more than the highest, who barely surpass the lowest scoring fifth-graders!

Support in the Literature

The loglinear analysis supports the findings of correlation between student predictions and actual adult ratings, as well as the theories of Gardner (Brandt, 1988; Hatch & Gardner, 1986; Gardner & Hatch, 1989) and Wolf (1989) that measures of judgment and self-awareness are needed to demonstrate ability development. They also converge with Flood and Lapp (1989), Della-Piana et al.(1988), Cooper (1990) and Bishop (1987, 1989) in stressing the need to attend to audience awareness and the match between student and teacher judgment.

In particular, the emergence of emotional agreement as a key predictor of growth from grade 5 to grade 8, and the apparent lack of growth in high school, supports the theories of Boy (1990), Eisner (1990, November), and Corey and Corey (1987). Boy argues that the inability to deal with the emotional content of experience often impedes our ability to exercise skills and cognitive strengths. In grade 8, both the language and emotional agreements increase with score. Boy also points out that, as a culture, we tend to focus on the conceptual content of experience due to a lack of emotional vocabulary. Consistently high agreement on conception confirms this assertion.

Boy also relates the inability to reflect client feelings with three common failures of counselors. First, such counselors tend to analyze immediately, explaining the meaning of client behavior. Second, the counselor often tends to jump to offer his or her solution to the client. Third, counselors who do not reflect client feelings frequently move too quickly toward the solution of a less important problem, without giving the client time enough to focus on another more serious problem.

Writers and their teachers with whom they confer may also be considered clients and counselors striving to effect change in human behavior. And the patterns of judgments and choices by both teachers and students indicate high school teachers may be subject to the failures listed above. First, in this high school sample, agreement on conception, or analytical features of the work, remains high, while agreement on the emotions drops. Second, several factors indicate high school teachers tend to define students' problems and solutions for them in classroom writing: the number of assigned class papers appearing in the sample, the

requests by raters for the original assignments, the differences in writers' choice of modes between the test and the portfolios, and the raters' preference for the personal flair of the test pieces. Finally, duration of work in the high school is so short as to indicate teachers assign quick turn-around time, failing to give their students time to warm up and address more serious problems they discover themselves.

Eisner (1990) and Corey and Corey (1987) find traditional "measurements" intended to assess growth inadequate because they omit key emotional and attitudinal factors from the "list." Both the bias and ineffectiveness of test writing in predicting classroom performance, as well as the prominence of emotional agreement above language and conceptual agreement as an indicator of growth support their calls for assessment that includes these aspects of human learning.

Chapter 6
Limitations, Implications and Conclusions

Participation

As a research project this study has had to obtain informed assent and consent from the students and their parents. Although approximately the expected percentage of participation (70%) developed in grades five and eight, many fewer high school students returned portfolios (46%). Overall, the use of volunteers may increase the number of successful students participating, but special factors seem to be operating at the high school level.

In many high schools students enroll in semester courses, so that the teacher who begins with them in the fall may not be the same teacher collecting their portfolios in the spring. Second, whereas a fifth grader's writing folder may well contain pieces from social studies, math, and science investigations, high school teachers seldom see writing from another classroom teacher's assignments. Finally, many elementary school portfolios arrived photocopied, including covers, artwork, and text, obviously in an attempt to preserve the treasured originals for the writers. High school papers, on the other hand, tended to be original drafts, complete with teacher comments, red marks, and corrections. How many more of these "dead-letter" pieces never survived long enough to be selected for the portfolio at all? Clearly, if we are to assess high school writing abilities by the use of portfolios, we need to make portfolios a part of high school writing programs.

Range

Range of modes of discourse included in these portfolios correlated less consistently with test or portfolio score than in the pilot study (Simmons, 1990). Since one rater sorted all pilot study papers into mode of discourse groups, but many classroom teachers performed this function in the current project, method of collection may have increased the error. In order to preserve access to this apparently enlightening measure of ability, some means must be found to reliably classify large numbers of papers into mode of discourse groups. Perhaps a random sample of all incoming papers could be re-classified by one rater, and any set failing to reach .7 interrater reliability could be totally re-classified.

Scope of the Investigation

The grade 5 and grade 8 samples were rated in about the same time formerly required to score prompted essays reliably. The eleventh grade readers actually finished early, but fewer papers and raters arrived for the high school sample. If school districts currently using scored, globally-administered writing samples wish to keep expenses about the same as they move to portfolios, they will need to content themselves with a sample size of under ten percent of the population.

In that case, districts will also need to limit the number of factors investigated in any one sample in order to address the error rate experimentwise. Robust findings, such as test/portfolio correlations and differences, length and duration differences across grades, and the correlation of student and teacher judgments, will probably be unaffected by samples this size.

However, more marginal findings, such as mode of discourse differences, effect of month of writing, differences among districts, and socio-economic status correlations, will need large sample sizes to be reliable.

Of course, more factors suggest themselves as a result of this research. Many non-statistical investigations, such as the ones in Vermont or New Jersey, examine progress from draft to draft within a portfolio. At a minimum, it would be interesting to investigate the relation between the number of drafts a paper required and the final rating. Also, the work of Pearson (1988) and Wolf (1989) suggests that peer comments, minimally reflected as a number or a set of categories of helpers, might illuminate composing abilities.

Given the number of studies requiring certain forms to be included and the concern of my raters and those of Martin (1988), the nature of the original assignment, and the degree to which student choice of paper matches teacher choice might also be included to determine the degree to which they correlate with other measures. Specifically, investigations might determine whether the paper was written in response to a teacher- or a self-assignment, and whether or not the classroom teacher agrees with the student's decision to include certain papers.

Finally, testing groups must decide the focus of investigation. The consortium for which I worked said they intended to evaluate programs, not individual students. Yet, when the results came in, many superintendents, principals and board members complained that they did not have feedback on specific students. The depth of data generated by portfolios will be much more expensive to gather from total populations, while the single-measure tests we have

used globally are both damaging and misleading.

General Conclusions

As a result of this research, I recommend three changes in traditional approaches to writing instruction and assessment. First, all of us involved in the development of writing abilities must work to stop the use of prompted essay tests as measures of writing ability for gatekeeping purposes. The current findings clearly indict these measures as inaccurate predictors of writing performance, as reflected in actual classroom papers chosen for the portfolios. Moreover, prompted performance measures negatively impact the weakest writers, while the stronger ones benefit. Since it also seems writers from the poorest schools suffer most, we must no longer allow expediency and tradition to dictate how we measure writing abilities, or the capacity of our institutions to develop it.

Second, educational leaders and classroom teachers must commit themselves to an overhaul of high school writing curricula. If high schools across this country are anything like the ones included in this study, and teachers with whom I have shared these results assure me that they are, high school students are routinely deprived of their voices and their self-determination as writers in order that writing may be used to transact the day's business. Teachers apparently give writing assignments providing such short duration and requiring so little personal commitment, that adult readers prefer reading prompted essays to reading selections of their colleagues' assignments. Moreover, these older students

seem at a loss to match adult judgments about their work, when eighth graders do it well.

Third, our educational community, facing the exhortations of a President and many governors to do otherwise, must heed the challenge of Broadfoot (1988) to opt for "descriptive" rather than "competitive" assessment. Competitive evaluations seek to fit all local variations to a national standard, following the tradition of the industrialized state. But our information economy needs, and writing abilities respond to, climates in which the workers and learners are invited into the evaluation process, and the workers' thinking processes themselves, including attitudes, values and emotions, become the focus of investigation, instead of more easily managed products.

References

Anastasi, A. (1982). Psychological Testing , Fifth Edition. New York: Macmillan Publishing Co., Inc.

Apple, M. (1986). Teachers and Texts:a political economy of class and gender roles in education. New York: Routledge & Kegan Paul.

Aristotle. (1984). The Rhetoric and the Poetics of Aristotle. Roberts, W.R., tr. New York: Modern Library.

Barba, M.P., Carrolton, E.T. & Yeaw, E.M.J. (1985). Portfolio Assessment: An Alternative Strategy for Placement of the RN STudent in a Baccalaureate Program. Innovative Higher Education, 9, 2, 121-127.

Benderson, A. (Ed.) (1989) Focus 23 The Student Writer: An Endangered Species? (Available from Educational Testing Service, Princeton, NJ)

Benedict, S. (1989). Looking at their Own Words: Students' Assessment of their Own Writing. Manuscript submitted for publication.

Berger, S., Dertouzos, M.L., Lester, R.K., Solow, R.M. & Thurow, L.C. (1989). Toward a New Industrial America. Scientific American. 260, 6, 39-47.

Bishop,W. (1987). Revising the Technical Writing Class: Peer Critiques, Self-Evaluation and Portfolio Grading. Paper presented at the Annual Meeting of the Penn State Conference on Rhetoric and Composition, 6th, State College, PA, July 7-10, 1987. (ERIC Document Reproduction Service ED 285 178).

Boy, A.V. (1990). The Reflective Process. In Boy, A.V. & Pine, G. J. A Person-Centered Foundation for Counseling and Psychotherapy (pp.23-54). Springfield,Il: Charles C. Thomas.

Brandt, R. (1988). On Assessment in the Arts: A Conversation with Howard Gardner. Educational Leadership, 45, 30-34.

Breland, Hunter M., Jones, Robert J. & Educational Testing Service (1984). Perceptions of Writing Skills. Written Communication, 1, 101-119.

Broadfoot, P. (1988). Profiles and Records of Achievement: a real alternative. Educational Psychology, 8, 4, 291-297.

Bruffee, K. A. (1988). On Not Listening in Order to Hear: Collaborative Learning and the Rewards of Classroom Research. Journal of Basic Writing, 7, 3-12.

Buddmeier, R.E. & Raivetz, M.J. (1990). '95 of Bust: Studying Writing in an Urban District as the Class of '95 Heads Toward a High Risk, Statewide Graduation Test. (ERIC Document Reproduction Service ED 318 765).

Burnett, D.G. (1985). Giving Credit Where Credit is Due: Evaluating Experiential Learning in the Liberal Arts. Innovative Higher Education, 10, 1, 43-54.

Business Council for Effective Literacy. (1990). Standardized Tests: Their Use and Misuse. BCEL Newsletter for the Business Community, 22, 1.

Cohen, M. (1990). Test Questions: a subject for the nineties. The Boston Sunday Globe (December 2, 1990) A33, A42.

Comstock, M. (1988). When Children Are Thinkers. Unpublished manuscript. University of New Hampshire.

Cooper, W. (1990). Planning the AVID Portfolio:Designing a Tool to Prepare College Bound Minorities. Portfolio News, 2, 1, 3, 13-14.

Corey, M.S. & Corey, G. (1987). Groups:Process and Practice, third edition. Pacific Grove, CA: Brooks/Cole Publishing Co.

Crowhurst, M. & Piche, G. L. (1979). Audience and Mode of Discourse Effects on Syntactic Complexity in Writing at Two Grade Levels. Research in the Teaching of English, 13, 101-109.

Dagavarian, D.A. (1989). Portfolio Assessment. (ERIC Document Reproduction Service ED 306 894).

Della-Piana, G.M. et al. (1988). What Assessment of Reader-Writer Conferencing Can Externalize. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA, April 5-9, 1988 (ERIC Document Reproduction Service ED 294 171).

Diederich, Paul (1974). Measuring Growth in English. Urbana, IL: NCTE.

Dixon, W.J. (Ed.) (1988). BMDP Statistical Software Manual. Berkeley, CA: University of California Press.

Eckhardt, C. D. & Stewart, D. H. (1981). Towards a Functional Taxonomy of Composition. In G. Tate & E. P. J. Corbett (Eds.), The Writing Teacher's Sourcebook (pp. 100-107). New York: Oxford University Press.

Eisner, E. (1990, November). Keynote speech presented at the Alternative Assessment Conference of the Concord (NH) School District and the N.H. Department of Education, Merrimack, NH, Nov. 8, 1990.

Elbow, P. & Belanoff, P. (1986a). Portfolios as a Substitute for Proficiency Examinations. College Composition and Communication, 37, 336-339.

Elbow, P. & Belanoff, P. (1986b). Using Portfolios to Increase Collaboration and Community in a Writing Program. WPA: Journal of Writing Program Administration, 9 (Spring, 1986).

Faigley, L., Cherry, R. D., Joliffe, D. A. & Skinner, A. M. (1985). Assessing Writers' Knowledge and Processes of Composing.

Norwood, NJ: Ablex Publishing Corp.

Flood, J. & Lapp, D. (1989). Reporting Reading Progress: A Comparison Portfolio for Parents. Reading Teacher, 42, 7, 508-514.

Flower, Linda (1981). Writer-Based Prose: A Cognitive Basis for Problems in Writing. In G. Tate & E. P. J. Corbett (Eds.), The Writing Teacher's Sourcebook (pp. 268-293). New York: Oxford University Press.

Freedman, Sarah Warshauer (1979). How Characteristics of Student Essays Influence Teachers' Evaluations. Journal of Educational Psychology, 71, 328-338.

Freedman, Sarah Warshauer (1983). Student Characteristics and Essay Test Writing Performance. Research in the Teaching of English,
17, 313-325.

Gardner, H. & Hatch, T. (1989). Multiple-Intelligences Go to School. Educational Researcher, 18, 4-10.

Geiger, J. & Shugarman, S. (1988). Portfolios and Cases Studies to Evaluate Teacher Education Students and Programs. Action in Teacher Education, 10, 3, 31-34.

Getzels, J.W. & Csikszentmihalyi, M. (1976). The Creative Vision: A Longitudinal Study of Problem Finding in Art. New York: John Wiley & Sons.

Glickman, C.D. (1990). Open Accountability for the '90s:Between the Pillars. Educational Leadership, 47, 7, 38-42.

Godshalk, F. I., Swineford, F. & Coffman, W. E. (1966). The Measurement of Writing Ability. New York: College Entrance Examination Board.

Graves. D. H. (1983). Writing:Teachers and Children at Work.

Portsmouth, N.H.: Heinemann Educational Books.

Graves, D. H. (1973). Sex Differences in Children's Writing. In Graves, D., A Researcher Learns to Write (pp. 7-15). Exeter, N.H.: Heinemann Educational Books, Inc.

Greenberg, K. & Witte, S. (1988). Validity Issues in Direct Writing Assessment. In Greenberg, K & Slaughter, G. (eds.). Notes from the National Testing Network. Volume VIII, November 1988. (ERIC Document Reproduction Service ED 301 888).

Hansen, J. (1987). When Writers Read. Portsmouth, N.H.: Heinemann.

Hatch, T.C. & Gardner, H. (1986). From Testing Intelligence to Assessing Competencies: A Pluralistic View of Intellect. Roeper Review, 8(3), 147-150.

Henderson, W.E., Jr. (1982). Articulated Instruction Objectives Guide for Drafting. Final Document for Articulation of Drafting. Greenville County School District, Greenville, SC (ERIC Document Reproduction Service ED 220 579)

Hyde, J. S. & Linn, M. C. (1988). Gender Differences in Verbal Ability: A Meta-Analysis. Psychological Bulletin, 104(1), 53-69.

Jesser, D.L. (1984). The Employability Skills Initiative in Colorado. Journal of Career Development,11, 1, 33-41.

John-Steiner, V. (1985). Notebooks of the Mind. New York: Harper and Row Publishers.

Kemp, D., Cooper, W. & Davies, J. (1990). The Role of Administration in Portfolio Development. Portfolio News, 2, 1, 1,12-13.

Killingsworth, M. & Sanders, S. (1987). Portfolios for Majors in Professional Communication. Technical Writing Teacher, 14, 2, 66-69.

Kinneavy, James L. (1971). A Theory of Discourse. Englewood Cliffs,

NJ: Prentice-Hall, Inc.

Knoke, D. & Burke, P.J. (1980). Log-linear models. Beverly Hills, CA: Sage Publications.

Lammon, K. R. (1985). Job Search Techniques for Fine Artists: An Advisor's Handbook. Paper presented at the Annual Meeting of the American College Personnel Association, Boston, MA, March 24-27, 1985. (ERIC Document Reproduction Service ED 260 352)

Lunsford, Andrea L. (1981). Cognitive Development and the Basic Writer. In G. Tate & E. P. J. Corbett (Eds.), The Writer's Sourcebook (pp. 257-267). New York: Oxford University Press.

Marsh, H.F. & Lasky, P.A. (1984). The Professional Portfolio: Documentation of Prior Learning. Nursing Outlook, 32, 5, 264-267.

Martin, W. (1988). Dancing on the Interface: Leadership and the Politics of Collaboration. Writing Program Administration, 11, 3, 29-40.

Marzano, R.J. & Costa, A.L. (1988). Question:Do Standardized Tests Measure General Cognitive Skills? Answer: No. Educational Leadership, 45, 8, 66-71.

Mathews, J.K. (1990). From Computer Management to Portfolio Assessment. Reading Teacher, 43, 6, 420-21.

McCready, M.A. & Melton, V.S. (1981). Feasibility of Assessing Writing Using Multiple Assessment Techniques. Research Report. Ruston: Louisiana Technical University (ERIC Document Reproduction Service ED 220 871)

McLean, L. (1987). Emerging with Honour from a Dilemma Inherent in the Validation of Educational Achievement Measures. Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC, April 20-24, 1987. (ERIC Document Reproduction Service ED 286 904)

Moffett, James (1981). Active Voice. Upper Montclair, NJ: Boynton/Cook Publishers, Inc.

New Hampshire Association of Teachers of English (1990). State Board Plans to Kick the CAT. NHATE Newsletter, 10, 2, 1.

Newkirk, T. (1984). How Students Read Student Papers. Written Communication, 1(3), 283-305.

Olson, Miles C. & Swadener, Marc (1984). Establishing and Implementing Colorado's Writing Assessment Program. English Education, 16, 208-219.

O'Neil, J. (1991). Education Goals Await Panel's New Monitoring Scheme. ASCD Update, 33, 3; 1, 6, 8

Pearson, H. (1988). The Assessment of Reading through Observation. Reading, 22, 3, 158-163.

Peters, E. (1981). Basic Skills Improvement Policy Implementation Guide #2 (Revised Edition) Writing Assessment Manual. Massachusetts Department of Education.

Piaget, J. & Inhelder, B. (1969). The Psychology of the Child. New York: Basic Books, Inc.

Rose, M. (1989). Lives on the Boundary: The Struggles and Achievements of America's Underprepared. New York: The Free Press.

Rosenblatt, L. (1978). The Reader, the Text, the Poem: the Transactional theory of the Literary Work. Carbondale, IL: Southern Illinois University Press.

Rothman, R. (1990). Large 'Faculty Meeting' Ushers in Pioneering Assessment in Vermont. Education Week,10, 6, 1.

Searle, D. & Stevenson, M. (1987). An Alternative Assessment Program in Language Arts. Language Arts, 64, 3, 278-84.

Sevigny, M.J. (1981). Triangulated Inquiry -- A Methodology for the Analysis of Classroom Instruction. Ethnography and Language in Educational Settings, Green, J. & Wallat, C. (eds.) Norwood, NJ: Ablex Publishing Corp.

Shannon, P. (1989). Broken Promises. Granby, MA: Bergin & Garvey Publishers, Inc.

Shrock,S.A. & Foshay, W.R. (1984). Measurement Issues in Certification. Performance and Instruction, 23, 1, 23-27.

Simmons, J. (1990). Portfolios as Large-Scale Assessment. Language Arts, 67, 262-268.

Spandel, V. & Stiggins, R. (1990). Creating Writers: Linking Assessment and Writing Instruction. New York: Longman.

Teale, William H. (1988). Developmentally Appropriate Assessment of Reading and Writing in the Early Childhood Classroom. The Elementary School Journal, 89, 172-183.

Terry, G.L. & Eade, G. E. (1983). The Portfolio Process: New Roles for Meeting Challenges in Professional Development. Paper presented at the 63rd Annual Conference of the Association of Teacher Educators, Pensacola, FL, Jan. 29-Feb. 4, 1983. (ERIC Document Reproduction Service ED 229 342).

Tabachnick, B.G. & Fidell, L.F. (1989). Using Multivariate Statistics, second edition. New York: Harper and Row.

Valencia, S., Pearson, P.D., Peters, C.W. & Wixson, K.K. (1989). Theory and Practice in Statewide Reading Assessment:Closing the Gap. Educational Leadership, 46, 7, 57-63.

Vygotsky, L. (1978). Mind in Society. Cambridge, MA: Harvard University Press.

Ware, E. (1988). Helping Students to Prepare a Technical Communications Portfolio. Technical Writing Teacher, 35, 1,

56-62.

White. E.M. (1988). Reliability Revisited:How Meaningful are Essay Scores? In Greenberg, K. & Slaughter, G. (eds.) Notes from the National Testing Network in Writing. Volume VIII, November 1988. (ERIC Document Reproduction Service ED 301 888).

White, E. M. (1985). Teaching and Assessing Writing. San Francisco: Jossey-Bass Publishers.

Williamson, R.E. & Abel, F. J. (1989). the Professional Portfolio:Keys to a Successful Job Search for the Beginning Teacher. Paper presented at the Annual Meeting of the Association of Teacher Educators, St. Louis, MO, Feb. 18-22, 1989. (ERIC Document Reproduction Service ED 304 418).

Wolf, D.P. (1989). Portfolio Assessment:Sampling Student Work. Educational Leadership, 46, 7, 35-39.

Wolfe, M. (1989). Rethinking Assessment: Issues to Consider. Information Update, 6, 1, 4-5.

Appendix A Instructions to Principals

Dear Principal:

I am writing to ask your help in administering the Writing Sample for 1989-90. As you may have heard, we will be examining portfolios of writing students select as their best, as well as rating the traditional timed writing test. We are able to evaluate more papers in greater depth by collecting the work of fewer students, that is, a randomly chosen subset of all students in the districts. We would like to avoid putting pressure on certain students, however, by not selecting them early in the year.

Accordingly, we will notify each school by March 15 of students from whom we would ask you to collect a portfolio of 3 pieces they have written during the school year, as well as a timed-test written sample collected only from the selected students. So that students will have writing to choose from, we need to ask that all fifth, eighth and eleventh grade teachers help their students save writing completed during the school year, beginning immediately.

For self-contained fifth grade classrooms using writing folders, this process is already in place. Eighth and eleventh grade English teachers who have the same students all year may simply provide in-school folders for students to store their completed drafts from English and other subjects as well. For schools in which students rotate between semester courses or where courses enroll students across grade levels, special arrangements will probably be needed. Perhaps students who switch teachers can take their folders with them. Please share with us solutions that you find work best for you.

I have already met with writing sample coordinators to discuss plans for collecting more detailed information from the portfolios. These specific instructions will follow with the names of randomly-selected students by March 15. We are enlisting the aid of all principals to insure that each building in all the SAU's gets the same directions at the same time.

For your information I have included a copy of a manuscript accepted by Educational Leadership outlining this new evaluation method and its rationale and benefits.

Thank you for your time and effort. If you have any questions, please contact ...

Appendix B Portfolio paper coversheet

Coversheet

Student Number _____ **Gender:** Male Female
Paper Number 1 2 3 **Grade:** 5 8 11 **Mode:** N D E A P
3 reasons this paper shows how good a writer you are: (Use back if needed)

_____
_____
_____
_____
_____
_____
_____
_____

Month you started or got the assignment (Circle one):

Sept	Oct	Nov	Dec	Jan	Feb	Mar
1	2	3	4	5	6	7

Number of days until you finished _____

Length of the paper in words _____

Compared to writing of <u>other people your age</u>, how would <u>you</u> rate this paper? (Circle one):

	2	3	4	5	6	7	8	
Among the worst								Among the best

Compared to writing of <u>other people your age</u>, how would you expect <u>a teacher</u> to rate this paper? (Circle one):

	2	3	4	5	6	7	8	
Among the worst								Among the best

Appendix C Instructions sent to classroom teachers relative to collection of student portfolios

Dear cooperating teacher:

Thanks for helping out in this first local assessment of writing ability through the use of portfolios with large numbers of students. We hope this year's process will be more meaningful to all involved: students, teachers and administrators.

Enclosed please find both **instructions** and **coversheets** to use in collecting portfolios and test samples from students in your classes who have been randomly selected and volunteered to take part in this year's study. Since we are reading more and longer papers from each student, fewer students will take part than have done so in the past. Please convey to your students that they are taking part in an important attempt find out more about student writing abilities, opinions and habits.

As you collect the portfolios with your students, please keep track of what works well for you and what does not work so well. Teachers created this procedure and your observations will improve it for the future.

When your principal tells you which of the randomly-selected students have volunteered and returned parental consent letters, please do the following things:

1) make arrangements for students to write the timed writing sample (one and a half hours)

2) when the samples are finished, quickly read over the ones from your students and decide the **mode of discourse** (narrative, description, exposition, argument, or poetry). Of course, many pieces of writing combine several modes, but choose the mode you feel best describes the **main purpose** of the piece: N = tells a story, D = describes a person, place or thing, E = explains an idea or process, A = convinces a reader to adopt and opinion or a course of action, P = written in a form intended to be different from prose

3) write N, D, E, A, or P next to the student number on the final draft sheet

4) ask your students to look back over the writing they have done this year **from any subject or course, including teacher-assigned or self-assigned pieces.** Ask them to choose three which show **how good a writer he or she is**

Appendix C, continued

5) start a **Coversheet** for each of the three pieces selected by the student. Fill in student number, student gender, paper number (simply call one paper "1", another "2" and the other one "3"), grade level, and circle N, D, E, A, or P to indicate the primary mode of discourse of the paper

6) ask the student to write 3 reasons for each paper that tell why the paper shows how good a writer the student is

7) ask the student to circle the month she/he started the first draft or received the assignment, **whichever is first.** The student should then count the number of calendar days that elapsed before the final draft was finished and write that number on the line provided. Notice, we are not looking for the number of days the student can recall actually putting pencil to paper, but the time elapsed

8) ask the student to count the number of words in the paper (including every word)

9) ask the student to circle the rating he or she would give the paper, compared to the writing of other students of the same age. Also have the student circle the rating he/she would expect a teacher to give the paper, when the teacher compares it to others written by students of the same age

10) when the students have finished the coversheets for all three of their papers, please collect the papers and the coversheets. Make sure each paper is identified only by the student number and the paper number (e.g. #3104-1, 2 or 3). Teachers who attended the planning sessions earlier this year asked that the portfolio papers be the cleanest copies possible. We do not intend that you or the student re-copy papers merely for the assessment. Please be sure to remove school, student or teacher identifying marks, however, and try to mask other comments where possible (clear photocopies are fine)

11) please return to your principal in separate stacks, the timed test papers (with mode of discourse indicated), the coversheets, and the portfolio papers (with student number and paper number indicated).

I know this is a lot to ask. I hope the random selection process leaves you with only a few students to handle. I also hope you find watching the students' selection process fascinating. Most of all, I thank you for your time and commitment to the teaching of writing.

If you have any questions, feel free to call us. We look forward to receiving your students' portfolios by April 2.